Predicting informativeness of words from human brain signals

Lauri Kangassalo

Masters thesis
UNIVERSITY OF HELSINKI
Department of Computer Science

Helsinki, January 21, 2019

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Lauri Kangassalo

Työn nimi — Arbetets titel — Title

Predicting informativeness of words from human brain signals

Oppiaine — Läroämne — Subject
Computer Science

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Masters thesis	January 21, 2019	49+12

Tiivistelmä — Referat — Abstract

We study the effect of word informativeness on brain activity associated with reading, i.e. whether the brain processes informative and uninformative words differently. Unlike most studies that investigate the relationship between language and the brain, we do not study linguistic constructs such as syntax or semantics, but informativeness, an attribute statistically computable from text. Here, informativeness is defined as the ability of a word to distinguish the topic to which it is related to. For instance, the word 'Gandhi' is better at distinguishing the topic of India from other topics than the word 'hot'. We utilize Electroencephalography (EEG) data recorded from subjects reading Wikipedia documents of various topics. We report two experiments: 1) a neurophysiological experiment investigating the neural correlates of informativeness and 2) a single-trial Event-Related brain Potential (ERP) classification experiment, in which we predict the word informativeness from brain signals. We show that word informativeness has a significant effect on the P200, P300, and P600 ERP-components. Furthermore, we demonstrate that word informativeness can be predicted from ERPs with a performance better than a random baseline using a Linear Discriminant Analysis (LDA) classifier. Additionally, we present a language model -based statistical model that allows the estimation of word informativeness from a corpus of text.

ACM Computing Classification System (CCS):
- Applied computing~Life and medical sciences
- Human-centered computing~Human computer interaction (HCI)
- Information systems~Language models

Avainsanat — Nyckelord — Keywords
EEG, ERP, BCI, Language models, Neurolinguistics, Machine learning, LDA

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

Abbreviations

The following abbreviations are used throughout the document.

AUC: Area Under the ROC Curve
BCI: Brain-Computer Interface
EEG: Electroencephalography
ERP: Event-Related brain Potential
ICA: Independent Component Analysis
LDA: Linear Discriminant Analysis
LM: Language Model
LMM: Linear Mixed Model
w.r.t.: with relation to

List of figures

List of tables

Appendices

1 Introduction

Language is a complex system of symbols that enables the efficient exchange of thoughts and experiences between humans. Although we use language daily, little is known of how language is represented internally by the brain. The relationship between the human brain and language is studied by a research field called neurolinguistics [55]. A majority of neurolinguistic research investigates how the brain processes syntax, semantics, and other linguistic constructs [37, 24, 44, 28]. Far less studied, however, have been the effects of the statistical attributes of language. Language statistics can reveal remarkable linguistic phenomena. Consider, for example, Zipf's law, which states that the most frequent word in a corpus occurs twice as many times as the second most frequent word, which in turn occurs three times as often as the third word and so on. This law holds with relatively high accuracy for all known human languages. A more recent example of the power of language statistics is presented by Piantadosi et al. [61] who, using a statistical model of language, show that a word's length is predicted by its information content. In other words, they demonstrate that longer words tend to have more information content than short words, and argue that this signifies that the human lexicon is structured for efficient communication. Furthermore, they demonstrate that this holds for 10 languages. Due to the universality of language statistics, an intriguing question is whether these statistics are somehow reflected in human cognitive processing, i.e. whether the brain captures the statistical nature of language. Taking this thought a bit further, it would be appealing to see whether these statistics of language could be "re-constructed" from brain signals alone. That is, by looking at the brain signals of someone reading text, could we decipher statistical attributes of the words they are perceiving?

Informativeness, or information content, is a concept employed in linguistics as well as information retrieval. For example in information retrieval, search results may be ranked according to document relevance with relation to a query term [67]. Intuitively, some words describe documents better than others. Consider, for instance, the words 'housecat' and 'is'. In addition to the intuition that 'housecat' is much more closely related to the topic of small felines than the word 'is', they differ in statistical terms: 'housecat' occurs more often in discourses about cats than other discourses, while the word 'is' occurs abundantly in any multi-sentence body of English text. Thus, 'housecat' can be considered to be more informative than 'is'. Here, word informativeness is defined via the entropy of a distribution of documents' probabilities for generating a given word. In other words, high entropy, or uncertainty of a document given a word, signifies that the word is uninformative.

In a typical neurolinguistic experiment, the stimuli (words, sentences, discourses etc.) are hand-crafted to produce a certain effect in the brain via syntactical or semantical anomalies (e.g. [18, 58, 21, 32]). While these studies

have provided valuable insights on the cognitive processes associated with language processing, the results may not be fully extrapolated to reading natural, everyday language. Contrary to these studies, the goal of this thesis is to investigate natural discourse reading. That is, reading multiple consecutive sentences which have not been specifically crafted for proving hypotheses nor which contain anomalies inserted on purpose. To achieve this goal, Electroencephalography (EEG) data recorded from participants reading English Wikipedia articles of various topics is analysed [16]. Furthermore, instead of studying syntax and semantics, which are human-created linguistic constructs, this thesis studies the effect of word informativeness, a statistical measure inherently present in any human language. To our knowledge, this is the first study to investigate the relationship between informativeness in natural language and brain activity associated with reading.

To study brain activity associated with words, this study focuses on Event-Related brain Potentials (ERPs), which are electrophysiological brain responses time-locked to stimuli. ERP studies are commonplace in neurolinguistics. For example, the ERPs associated with each word in a sentence could be compared with each other to study the differences in cognitive processing of differing words. Moreover, ERPs can be classified using machine learning methods, a prominent approach especially in the Brain-Computer Interface (BCI) community [7, 47]. The ERPs are classified on a single trial basis, meaning that each ERP is considered a 'trial', and these trials are classified separately from each other. For instance, given ERPs matching to words, the words could be categorised by classifying each of the ERPs. In order to improve the signal-to-noise ratio of brain signals in BCI experiments, the subjects are typically instructed to imagine some motor activity or otherwise control their brain activity [74]. Analysing this type of brain activity, however, provides little insight to the brain processes associated with the task itself, and typically requires some training on the user's part.

In this study, we are interested in the natural brain responses of reading. Passive BCI is a relatively new alignment in the BCI community [79], which provides brain-computer interaction without the user's conscious efforts to control their brain activity. In the EEG data utilized in this thesis, the participants were instructed to not engage in any additional mental or other activity while performing the reading task, so that the brain signals measured would reflect the cognitive processes related to reading. The focus of this thesis is to predict the informativeness of words based on these "passive" brain signals. These types of predictions could be used in passive BCI systems to enhance the interaction between man and machine.

To summarize, this work studies the effects of word informativeness on brain activity associated with reading natural language text, and aims to predict the word informativeness from the brain signals of the reading participants. Consequently, the hypotheses of this thesis are:

H1: Brain activity associated with reading natural language text is affected by the informativeness of the read words.

H2: The informativeness of words can be predicted in a single-trial scenario from the brain activity associated with reading the words.

The experimental results reported in this work indicate that the informativeness of words has a significant effect on brain activity associated with reading, and demonstrate that prediction of word informativeness from ERPs is possible.

The structure of this thesis is as follows. The background chapter presents the characteristics of EEG data, and the state of the art of neurolinguistics and brain-computer interfaces. Chapter 3 defines word informativeness and presents a model for estimating the informativeness from a corpus of text. Chapters 4 and 5 discuss the methods and results of the neurophysiological experiment on and the single-trial prediction of the word informativeness. These chapters are concerned with the research questions H1 and H2, respectively. The final chapter presents conclusions based on the previous chapters and provides further discussion.

2 Background

The relationship between language and the brain can be studied with brain imaging technologies, such as functional Magnetic Resonance Imaging (fMRI), Magnetoencephalography (MEG) and Electroencephalography (EEG). EEG was used in the experiment discussed in this thesis, so EEG will be the main focus of the background section.

2.1 Electroencephalography

EEG measures the voltage fluctuations produced by neurons in the brain from the surface of the scalp. Modern EEG-equipment consists of a signal amplifier and a cap that contains the electrodes that measure the voltage fluctuations from various locations on the scalp. Unlike many other brain imaging methods, EEG equipment is relatively cheap and mobile. MEG for example requires shielding from external magnetic signals, such as Earth's magnetic field, which is typically achieved by building a specific room for the imaging equipment with the appropriate shielding. In addition to cost-efficiency and mobility, EEG is one of the oldest brain imaging technologies. Due to these factors there exists an abundance of literature on experiments conducted with EEG.

The invention of electroencephalography dates back to the end of the 19th century, when the English physician Richard Caton and the Polish physiologist Adolf Beck published their findings on electrical phenomena occurring at animal cortices, independently of each other [12, 27]. In these experiments, the measurements were taken directly from the cortices of

rabbits, cats, and dogs. A German physiologist and psychiatrist Hans Berger was the first to record electrical activity of the human brain in 1924 [27]. Unlike Caton and Beck, he measured the electrical activity in a non-invasive fashion from the scalp, naming his invention *das Elektrenkephalogramm* (electroencephalogram). EEG was further developed by Edgar Douglas Adrian and B. H. C. Matthews in 1930s [1], and later, in the 1950s, by William Grey Walter, who invented EEG topographies [6]. Since then, EEG has been used clinically to diagnose epilepsy [66], sleep disorders [59], and coma [78] amongst others. In addition, EEG has been used extensively to study the human cognition, with areas ranging from altered states of consciousness (induced by e.g. meditation [3]) to problem solving [19], and, of course, language [37, 24, 44, 28].

The field of neurolinguistics can be dated back to the latter half of the 19th century, when Paul Broca and Karl Wernicke published their independent studies on aphasia [15], a condition in which a person is unable to produce or understand speech, and which is caused by a physical damage to certain brain regions. At the time the name of this field was aphasiology, and the term neurolinguistics was not coined until the 1950s. Neurolinguistics gathered more and more interest in the latter half of the 20th century, with the improvement of brain imaging technologies such as EEG and MEG. These technologies have a high temporal resolution, with brain signals recorded at the millisecond range. This enabled studying the effects of single sentences and words on brain signals with *event-related brain potentials*.

2.2 Event-related potentials

EEG waveforms time-locked to stimuli are called Event-Related brain Potentials (ERPs) [49]. In a typical neurolinguistic experiment, the test subject is presented with some *stimuli*. The stimuli may be presented to the subject via various *sensory modalities*. Modality refers to the encoding in which information is perceived by humans. Sensory modalities include visual, auditory, tactile, olfactory, gustatory, and kinesthetic, of which visual and auditory are utilized in neurolinguistic studies. In the case of visual modality, the stimuli consists of random strings of letters, words, sentences, or other units of language, which are displayed to the subject. In other words, ERPs depict the changes in recorded scalp voltages correlated with some experimental event (e.g. a word displayed to the subject). To inspect the ERPs, the continuous EEG recording can be split to time-windows coinciding with the stimuli. That is, for each stimulus, take a small slice of EEG data containing the ERP associated with the stimulus. These slices are called *epochs*. Typically, epochs span 100-250 ms pre-stimulus and 500-1200 ms post-stimulus, depending on the experiment. Figure 1 illustrates epoching of raw EEG data.

ERPs consist of positive and negative voltage peaks. These voltage peaks are referred to as ERP-components, and they are linked to underlying

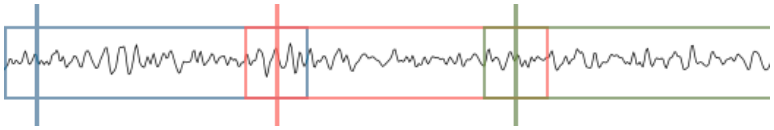


Figure 1: Raw EEG signal from one electrode segmented to epochs (rectangles) according to stimuli onset times (vertical bars), with time on the horizontal axis. The epochs and their matching stimulus onsets are color-coded. Note that in this case the epochs overlap, i.e. an epoch may contain parts of the previous/next ERP.

cognitive processes. ERP-components are named with the polarity of the peak as letter N or P for negative and positive followed by the latency of the peak in milliseconds post-stimulus. For example, the N170 ERP-component, which is associated with facial recognition, has a negative polarity and peaks 170 ms post-stimulus. However, since the ERP depicts continuous voltage fluctuations, the ERP-components may be mixed. Consider, for example, the N100 and P200 components, which are related to sensory processing of stimuli. A strong N100 component will "leak" to the following P200 component: the negative voltage produced by the cognitive event causing the N100 persists when the cognitive event causing P200 occurs, reducing the amplitude of the P200. Thus, when analysing an ERP-component, the voltage peaks occurring before it should be considered instead of only looking at the peak amplitude.

There are two types of ERP-components: exogenous and endogenous. Exogenous components are affected by the modality and physical characteristics of the stimulus, such as the amount of light entering the retina, and they are thought to index processing of sensory stimuli. N100, for example, is an exogenous component. Endogenous components are associated with higher cognitive processing, and typically are not affected by the input modality. P300 and N400 are examples of endogenous components. Typically, endogenous components occur later than exogenous components. Endogenous components will be the main focus in the rest of this chapter, since the focus of this thesis is in the cognitive processing of language.

Next, an overview of neurolinguistically relevant ERP-components and the cognitive processes associated with them will be given. For information on other types of ERP-components and their interpretations, see [34].

2.2.1 P300-component

In the 1950's and 1960's there was a trend of applying Shannon's information theory to experimental psychology [5]. Psychologists of the time were intrigued by the possibility that concepts formulated by Shannon, such as information and encoding, could be applied not only to electronic communication channels but the human cognition as well. In 1965, likely influenced by these thoughts, Sutton et al. [68] discovered the P300-component in a study

investigating the effects of stimulus uncertainty. In Sutton's experiment, the subjects were presented with a series of flashes of light and clicks while their ERPs to these stimuli were recorded. Each of the stimuli was preceded by a cue. In one of two experimental conditions, the stimuli following the cue was always a click or always a flash. In the other condition, the cues were followed by either a click or a flash. Between the stimuli, the subjects had to guess which of the two types of stimuli would come next. Sutton et al. found a "large positive deflection" occurring at roughly 300ms post-stimulus in the cases where the type of the upcoming stimulus could not be predicted (i.e. the latter experimental condition), and that the amplitude of this deflection varied with the stimulus probability. Sutton et al. named the deflection Late Positive Component (LPC), a name that is still used of the P300-component.

Subsequent studies have proposed multiple theories of how cognitive processes affect the amplitude and latency of P300. These studies argue that the amplitude of P300 correlates with context updating and allocation of attention, and that its latency is associated with task difficulty and individual cognitive capability [62]. Context-updating theory posits that the P300 indexes the updating of a mental representation of the stimulus: the underlying mental model has to be revised upon observation of an unanticipated stimulus. This updating requires attention, and it has been shown that P300 indexes attention allocation [33]: the P300 amplitudes tend to be smaller if more attentional resources are engaged by other tasks not related to the task whose ERPs are under scrutiny, than in cases where the attentional system is not taxed. Task relevance also plays a role: passive stimulus processing produces smaller P300 amplitudes than stimuli that are attended to. The latency of P300, on the other hand, is sensitive to the difficulty of the task: the harder the stimulus is to classify, the longer it takes to evaluate it, and the longer the latency of P300. The peak latency of P300 is quite volatile as it may vary between 250 - 500 ms or longer depending on the task. Furthermore, the latency is shorter in persons with superior cognitive performance and lengthens with normal aging and in some diseases, such as dementia.

Besides these general theories of the P300, the component has been studied with relation to language: in 1975, Friedman et al. [22] showed that the P300 occurs when language stimuli is processed. They displayed sentences word at a time to subjects and noted that each word elicited a P300. Furthermore, words that delivered task-related information elicited P300s with longer latencies than other words, and that the last word of a sentence produced a P300 of greater amplitude than any of the other words in the sentence. The latter was attributed to syntactic closure: the brain recognizes the sentence as a language unit when it ends, a process that manifests as a P300.

2.2.2 N400-component

Inspired by the study of Friedman et al. [22], Kutas and Hillyard tried to elicit P300s with semantically incongruous sentences [40]. Instead of evoking P300s, they stumbled upon the first language-specific ERP-component, the N400. In their 1980 published study [38], the subjects were presented with sentences that ended in a meaningful and expected way (e.g. "I take my coffee with milk and sugar"), and sentences that ended with an inappropriate and unanticipated word (e.g. "He took a sip from the transmitter"). They expected to see a P300 at the end of each sentence as per the effect of syntactic closure, but instead noticed a negative-going wave starting at 250 ms and peaking at 400 ms in the semantically incongruous sentences. They named this component the N400.

Further studies have shown that the N400 is highly context sensitive [37]. The amplitude of N400 has been shown to inversely correlate with the cloze probability of a word, which is a measure of the expectancy of a word. In other words, the less expected a word is in a context, the greater its N400 amplitude compared to more expected words [39]. For example, the last word in the sentence "he went to work by car" has a high cloze probability, while the last word in "he went to work by helicopter" has a low cloze probability, while both of them are semantically congruent. Thus, the latter sentence would produce a greater N400. Furthermore, N400 amplitude is affected by prior world knowledge: sentences which are in conflict with what the subject knows elicit N400s of greater amplitude [29]. For instance, in persons familiar with the Helsinki public transportation, the sentence "the metro trains in Helsinki are blue" will elicit a larger N400 than the sentence: "the metro trains in Helsinki are orange". This is due to the fact that the former sentence contradicts with such persons' world knowledge, as metro trains in Helsinki are, at the time of this writing, orange.

While the amplitude of N400 is affected by a multitude of factors, its latency is remarkably constant regardless of factors that commonly affect the latencies of other ERP-components; such factors include input modality, reaction times, and the difficulty of the task performed [37]. Furthermore, recent studies have found that N400 is not language-specific, but can be elicited with a variety of non-linguistic means, such as incorrect solutions to mathematical problems [37], incongruous endings of mute short films [37], and even cultural norm violations [51].

All of the aforementioned combined have led to speculation about N400 indexing a long-term multimodal memory access: N400 is thought to be a manifestation of the process of unifying the incoming stimulus with concepts stored in the long term-memory, regardless of the modality of the stimulus, and its amplitude to index the difficulty of the unification process [37].

2.2.3 P600-component

The P600-component was first reported by Osterhout and Holcomb in 1992 [58]. They noticed a positive voltage deflection occurring at around 500 to 800 ms post-stimuli in ungrammatical sentences. Namely, they displayed grammatical ("the woman struggled to prepare the meal") and ungrammatical ("the woman persuaded to answer the door") sentences to subjects, and found the P600 elicited by the word 'to'. It was also found that P600s could be elicited robustly with so called garden-path sentences, which are grammatical sentences that start in such a way that the reader's interpretation of the sentence will be most likely incorrect. For example, the sentence "the complex houses married and single soldiers and their families" leads the reader to first interpret that the sentence discusses complex houses, but since houses do not marry, the word 'houses' needs to be re-interpreted as a verb instead of a noun, rendering the sentence sensible. This led to the hypotheses that P600 indexes reanalysis and repair of obscure sentences, as well as syntactical integration of sentences [24, 26].

There is, however, controversy regarding the syntactical nature of processes eliciting P600. It has been suggested that instead of indexing syntactical processing, it reflects the broader task of updating a mental representation with new information [11]. Thus, garden-path sentences for instance require revising the mental model of the situation that was suggested by the beginning of the sentence. The updating of mental representations is reminiscent of what the P300 is thought to index. As it turns out, the P600 shares other traits with the P300 as well, and it has even been argued that P600 is, in fact, a late-occurring P300 [65].

2.2.4 Summary of ERP-components

To summarize, the P300, N400, and P600 have all been linked to language processing, and a variety of theories try to explain what cognitive processes they index. The P300 has been found to index uncertainty of upcoming events, giving rise to the context-updating theory [62]. The peak amplitude and latency of the P300 are affected by a variety of factors such as task difficulty and differences among subjects [62]. The N400 seems to index long-term memory access, and its latency is rather constant while its amplitude is affected by the task at hand [37]. The P600-component is perhaps the most debated. It has been associated with both syntactic and semantic processing, and according to one theory, it might be a late-occurring P300 due to the characteristics it shares with the P300 [65].

From the studies discussed in this chapter, it is evident that much of the literature on neurolinguistics is focused on studies where semantically or syntactically erroneous words, sentences, and discourses are presented to the participants. The effects of natural language are far less studied.

2.3 EEG data characteristics

EEG is a continuous recording of voltages from multiple electrode sites on the scalp. Thus, it has both temporal and spatial features. The temporal as well as the spatial resolution of the recording varies according to the experimental setup and the equipment used.

The number of electrodes vary across configurations, but their locations and labels are standardised [36]. Figure 2 (left) displays the locations and labels of 32 electrodes plotted on a head. The location of an electrode can be deciphered from its label: the letter in the label signifies the brain area over which the electrode is placed on the scalp: Pre-frontal (Fp), Frontal (F), Temporal (T), Parietal (P), Occipital (O), and Central (C). The electrodes residing between these have labels that combine these letters: FC, FT, CP, and TP. The number at each label corresponds with the electrode's lateral location: odd numbers for electrodes on the left hemisphere and even numbers for electrodes on the right hemisphere. The letter 'z' is reserved for electrodes at the lateral center. The brain areas after which the electrodes are labelled are shown in figure 2 (right).

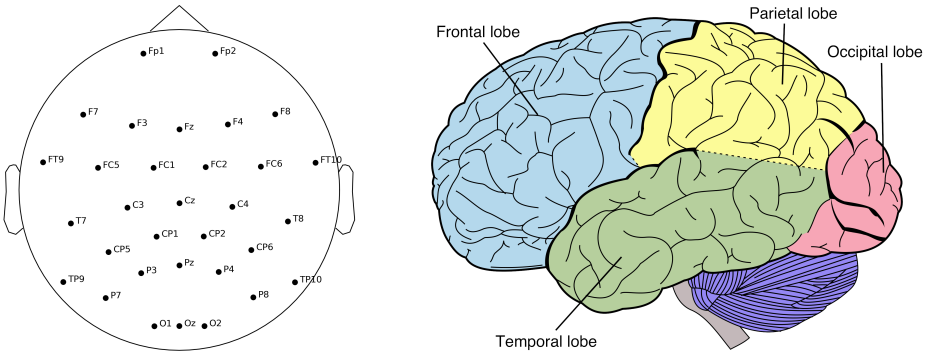


Figure 2: Left: the EEG montage used in the experiment; right: brain regions relevant for EEG.

Although the electrodes are labelled and placed according to the underlying brain regions, the spatial resolution of EEG is worse than in some other brain imaging techniques, such as MEG or fMRI. The human head forms a volume conductor which causes current to spread as the brain signal travels from its source, the neurons and synapses, to the electrodes recording it. This medium is inhomogeneous with varying conductivities in brain tissue, cerebrospinal fluid, the skull, and the scalp, which makes localising the signal source a difficult task given only scalp potentials; signal fluctuations in a given electrode may not always imply activity in the brain region directly under it [54]. Furthermore, the signals recorded by a single electrode are the averages of electric potentials generated by roughly 100 million to 1 billion neurons, which makes assigning activity to small groups of neurons infeasible.

Consequently, one has to be very careful in making assumptions about the underlying brain activity based on EEG signal.

The electrodes in EEG equipment are connected to a differential amplifier, which measures the difference between two inputs before amplifying the resulting signal. The two most common methods for computing this difference are common reference derivation and average reference derivation. In common reference derivation, the difference between an electrode input and a reference electrode input is computed. Typically, these reference electrodes are placed near the ears, but sometimes an electrode placed on the nose or on an even more distant body part. In average reference derivation, the average voltage of all of the electrodes is subtracted from the input of the electrode in question. Thus, the voltages at each electrode are always relative to some reference.

Despite the low spatial resolution of EEG, its temporal resolution is high. EEG records voltage fluctuations on the millisecond range. This is ideal for neurolinguistics, since language is processed rapidly by the brain. Typical sampling rates provided by EEG equipment vary between 240 and 2000 Hz, which correspond to 240 and 2000 data points (voltage measurements) recorded per second. The sampling rate defines the maximum possible frequency recorded according to the Nyquist criterion. The criterion states that for a sampling rate of r, the maximum frequency of the signal that can be interpreted from the samples is $0.5r$. So, for example if the EEG is recorded at 1000 Hz, the theoretical maximum frequency in the recording is 500 Hz. Neural oscillations occur at frequencies from 0.5 Hz to 100 Hz, and different frequency bands have been associated with different mental states. For example, alpha waves are neural oscillations at the frequency band 8-12 Hz, and they have been associated with wakeful relaxation, especially with eyes closed [3]. Other frequency bands have been associated with deep sleep (delta, 0.5-4 Hz) [2], and active concentration (beta, 12.5-30 Hz) [14], while the functions of some, such as the gamma wave at 30-100 Hz, remain disputed [71].

2.4 Artefacts and data cleaning

EEG equipment is designed to measure the voltage fluctuations produced by neurons. The amplitudes of these fluctuations are minuscule compared to electrical activity produced by other sources, and thus the signal of interest is confounded with noise. In order to improve the signal-to-noise ratio of EEG recordings and avoid reaching incorrect conclusions by analysing noisy data, it is important to know what the origins of the noise are and how the noise can be isolated from signal. This chapter introduces the common noise sources and basic data cleaning procedures for EEG.

2.4.1 Artefacts

The electrical activity that is not of a neural origin is termed artefact. Artefacts can be classified to physiologic and extraphysiologic artefacts: physiologic artefacts are generated by the test subject, but are not of a neural origin (e.g. changes in conductivity caused by sweating); extraphysiologic artefacts originate outside the test subject's body (e.g. electrical fields produced by equipment) [70].

Physiologic artefacts Muscle activity causes artefacts in the EEG, as muscle cells generate electric potentials as they contract. Typically muscle activity artefacts are caused by jaw clenching. Head movements also generate artefacts in the EEG data.

The most common artefacts in EEG data are caused by eye movements and blinks. An eyeball forms a dipole, with a negative pole oriented towards the retina and positive pole oriented towards the cornea. When this globe rotates, it generates an electrical field, which can be detected in the electrodes near the eyes. Eye blinks cause vertical eye movements, which cause artefacts that are most prominent in the pre-frontal electrodes Fp1 and Fp2. Horizontal eye movements, caused by e.g. reading, affect especially the lateral frontal electrodes F7 and F8. Since, however, the sum of electrical activity recorded by electrodes has to sum to zero, these artefacts affect the data of all channels. During an EEG experiment, eye movements are typically measured with Electrooculography (EOG) and included among the EEG data to aid with data cleaning.

Sweating changes the conductivity of the scalp, causing slow drifts of voltage. In addition to perspiration, respiration causes periodical movement of the head, which may also cause slow waves in the EEG recordings. Pulse may also generate artefacts, if an electrode is placed on a blood vessel. Naturally, these artefacts are in sync with the subject's pulse.

Extraphysiologic artefacts The electrodes may cause artefacts due to a poorly fitting cap, which does not keep the electrodes still. An electrode that is apart from the scalp or that is not stationary will cause the voltage measured by it to be flat or wildly fluctuating. The artefacts caused by electrodes are easy to distinguish from brain signal, since they usually occur only at one electrode site. Typically, less noise is present at the electrodes near the top of the head. This is because such electrodes generally remain more stationary than those residing on the sides, front, and back of the head.

The alternating current power grid generates a distinct electrical activity at 50-60Hz, depending on country. This type of artefact can be easily removed by filtering the recording appropriately.

2.4.2 Data cleaning

As discussed above, the EEG recording is teeming with artefacts from a variety of sources and with different attributes: some of the artefacts are more or less constant (e.g. power line noise), while others are transient (e.g. eye blinks), some are of a high frequency while others cause slow drifts in the recording spanning seconds or even minutes. It is recommended to aim to prevent these artefacts from occurring in the first place. For example, artefacts caused by eye blinks and voluntary movements of the participant can be reduced by instructing the participants to avoid excessive blinking and unnecessary movement during the experiment. Of course, not all of the artefacts can be eliminated this way; the subject can not stop breathing and the experimental equipment requires electricity to function. Thus, to improve the signal-to-noise ratio of EEG, the recording needs to be preprocessed prior to data analysis [48].

Typically, the first step in EEG preprocessing is filtering. The EEG data is filtered to include only the frequency bands of interest. Low-pass filtering the data below 50 Hz efficiently removes power line noise and other high-frequency artefacts, but is a poor choice if the brain activity of interest occurs at higher frequencies. In these cases, a notch filter may be used, which filters out only a small slice of bands near the power line frequency. Fortunately, in ERP studies the neural oscillations of interest (e.g. ERP components) occur at frequencies lower than 50 Hz, so the higher frequencies may be safely filtered out. High-pass filtering at low frequencies removes slow drifts in the recording caused by e.g. breathing. A recent study has shown, however, that high-pass filters with cutoffs at 0.3 Hz or higher produce artefacts in ERP studies, which may lead to incorrect conclusions of which ERP components were affected by the experimental conditions [69].

The measured voltages might drift over time due to e.g. sweating. To reduce the effect of slow drifts in voltages, the epochs can be baseline corrected. In ERP studies, the typical baseline correction procedure is the absolute baseline correction, where the average voltage of some small (say 200ms) pre-stimulus period of an epoch is computed and subtracted from all of the data points in the epoch. When this procedure is applied to all epochs in a recording, this leads to the epochs having similar average voltages. This approach assumes that the pre-stimulus periods of each epoch are unaffected by the experimental conditions.

In addition to filtering and baseline correcting the data, noise can be reduced by simply removing noisy channels and epochs from the data, a procedure called artefact rejection. Noisy electrodes produce voltage fluctuations of such a high amplitude that they can be picked from the EEG signal by visual inspection. On the other end of the spectrum, completely loose or malfunctioning electrodes produce a flat signal. The experimenter may inspect the EEG signals as they are being recorded from a subject or after the experiment in an off-line fashion, and mark them for rejection. In

addition to visual inspection of the EEG recordings, simple heuristics can be designed to remove noisy channels. Noisy channels can be rejected if the maximum and minimum voltages recorded by the corresponding electrode constantly surpass some threshold. On the other hand, channels with a flat signal can be rejected if the variance of the recorded voltages is sufficiently small. Noisy epochs may be rejected in the same fashion as noisy channels.

Epochs that contain voltage fluctuations above some threshold can be safely discarded, since the brain's electric potentials have amplitudes below such threshold. This is an efficient way of reducing transient noise, such as caused by eye blinks or movements of the head. Epochs and channels that are rejected from the data due to noise are commonly called bad epochs/channels.

Defining the rejection thresholds for channels and epochs can be tricky, since the amplitudes of brain potentials vary individually, due to for example skull thickness and shape. One approach to setting the threshold is defining a static percentage of epochs that have to be rejected. This is based on the notions that humans blink with an interval of 2 to 10 seconds, and eye blinks are the most common artefacts in ERP studies. As an example, consider an ERP study where the stimuli interval is one second. In such a setting, it is reasonable to assume that at least every tenth epoch contains an eye blink. Thus, 10% of the epochs with the highest absolute voltage values could be rejected on the assumption that they are contaminated by a blink. This percentage can then be adjusted to either direction depending on the noisiness of the data. However, when adjusting the threshold it is worthwhile to consider the trade-off between preserving the signal of interest and removing noise; a low threshold leads to loss of relevant signal whereas a high threshold increases the amount of noise present.

In addition to the aforementioned, rather straightforward techniques for EEG data cleaning, more advanced methods exist. One of the most popular ones consists of using Independent Component Analysis (ICA) to split the EEG signal to independent components, identifying the components that produce noise, and zeroing them out from the data [50]. This is an efficient way of correcting artefacts without losing precious data. The noise components can be identified by inspecting them with relation to stimulus onsets, looking at their power spectrum, and examining their scalp topographies. The voltage fluctuations of noise components are typically uncorrelated with stimuli onsets, their power spectrum does not match commonly known bands of neural oscillations, or their topographies suggest that they are of a non-neural origin, e.g. artefacts caused by eye movements have a frontal topography. The downside of ICA is that it requires manual analysis of the components and leaves a lot of room for interpretation for the analyser. ICA changes the waveforms of EEG radically, and so the analyser has to be especially careful when analysing the components and avoid removing any components containing signal of a neural origin.

2.4.3 Summary of artefacts and data cleaning

Multiple noise sources generate artefacts to the EEG recordings. Physiologic artefacts are caused by the subject, and include eye blinks and movements, muscle activity, and sweating amongst others. Extraphysiologic artefacts are caused by sources outside of the subject's body. Examples of these artefacts are poorly fitting electrodes and line noise generated by the AC power grid. It is clear that not all of the noise sources can be taken in to account when processing EEG data: a completely artefact-free dataset does not exist. However, with proper preprocessing it is possible to increase the signal-to-noise ratio of EEG data while preventing excessive loss of signals of interest.

Furthermore, it is to be stressed that EEG preprocessing can be conducted in many ways. Different EEG analysis require different preprocessing strategies and the analysers will always have some freedom in choosing the best methods and their parameters for the analysis in the particular case.

2.5 Brain-computer interfaces

In addition to studying the human cognition, brain activity recorded by EEG is used for human-computer interaction. The applications that enable this are called Brain-Computer Interfaces (BCI), a term that was coined in the 1970s when the first studies on BCIs were published by Jacques Vidal [72, 73]. Traditionally, BCIs have been used in a medical context to enable mobility and communication for patients with limiting conditions, such as paralysis [42]. BCI applications designed for healthy users are rather scarce, which is mostly because the input speeds of BCI systems are slow compared to the more traditional input methods, such as the keyboard and mouse. For example, the Hex-O-Spell mental typewriter, which enables the user to type text by controlling their brain signals achieves input rates from 2.3 to 7.6 characters/minute [8], which is much lower than the typing speeds on a regular QWERTY-keyboard. Recently, however, there has been more interest in using BCI with healthy individuals in the form of passive BCI.

2.5.1 BCI types

According to a categorization by Zander and Kothe [79], there are three types of BCI systems: active, reactive, and passive.

Active An active BCI, such as the Hex-O-Spell, requires the user to consciously control their brain activity to interact with a computer. For instance, the Hex-O-Spell is controlled by motor imagery. The user imagines right hand movements to move a pointer on a screen where characters are displayed and imagines right foot movements to select a character. Motor imagery alters sensimotor rhythms, which are oscillations recorded over sensimotor cortices. These oscillations can be recorded with EEG and the changes in them utilized to control a BCI system [75].

Reactive A reactive BCI is controlled with brain activity associated with some stimulus and modulated by the user. The P300 speller [17], another mental typewriting application, is an example of a reactive BCI. In the P300-speller study, the user was displayed with a 6x6 matrix consisting of letters and other symbols. The rows and columns of the matrix were flashed while the subject attended to a target symbol. Since the P300 is produced by presentation of a target stimulus among non-target stimuli, the target symbol could be deciphered based on the P300 response of the subject. Thus, the subjects modulated their brain activity by attending to a certain character, strengthening their P300 response to it.

Passive In passive BCIs, the users need not control their brain activity consciously. Due to this, the users are free to perform other tasks not related to the BCI control, and so passive BCIs can be used in combination with other input modalities. Furthermore, active and reactive BCI typically requires the user to learn to control their brain signals in order to interact with the computer, but since passive BCI only uses "passive" brain signals that occur without the user's explicit control, little training is required to use the system. An example of a passive BCI is the AlphaWoW, which was introduced in a study by Laar et al [43]. In AlphaWow one aspect of the player's character in the video game World of Warcraft (WoW) is controlled by changes in the alpha band of brain waves. In WoW the players control a character whose abilities are defined by its race and type. The objective of the game is to battle monsters or other players to evolve the player's character and acquire better equipment such as weapons and armour for the character. In the experiment, the players' character was a druid of the Night Elf race, which has the ability to turn itself in to a bear shape. In its druid shape, the character is vulnerable to attacks, but able to cast spells from a distance. In the bear shape the character can stand more damage done to it and is more efficient in close combat than in the druid form. The increase in alpha band activity, an event linked to relaxed alertness [3], was mapped to the druid shape, and the decrease in alpha band activity to the more aggressive bear shape. The shape of the character adapted based on the monitored changes in the alpha band, while the player used a keyboard and a mouse to control other aspects of the game. Thus, in AlphaWoW the uncontrolled brain activity of the players was used to enhance the human-computer interaction provided by other input methods.

2.5.2 Brain activity classification in BCIs

In order for a BCI system to respond to ongoing neurophysiological events, the brain signals need to be classified in a single-trial fashion. Single-trial classification means simply classifying samples one at a time. Consider, for example, the P300 speller, where the task is to classify the ERPs that contain a large P300 potential due to a perceived target stimulus. To solve this

problem with single-trial classification, each of the ERPs is given a label which indicates whether or not they were produced by a target stimulus. Then, a classifier is trained with this data and used to classify previously unseen ERPs one at a time. Support Vector Machines (SVMs) using kernel methods are popular non-linear classifiers in the BCI community while Linear Discriminant Analysis (LDA) is a good candidate amongst the linear classifiers [9]. While even neural networks have been used to classify brain signals [41], in ERP studies the classification problem is typically binary and solvable with a linear classifier [47]. Furthermore, simple classifiers such as SVMs and LDA are easier to understand than more complex methods such as neural networks, and by analysing the features they have learned from the recorded brain signals it is possible to gain insights on the underlying neurophysiological phenomena [31]. Of these classifiers, a binary LDA classifier for ERP data is described in the following chapter.

2.5.3 Linear discriminant analysis for classifying ERPs

Before using LDA to classify ERP data, the epochs containing the ERPs have to be transformed to a vector representation [9]. Depending on the application, the resulting feature vector may be composed of spatial features (voltages measured by multiple electrodes at a given time point relative to stimulus onset), temporal features (voltages measured by a single electrode at multiple time points relative to stimulus onset), or a mix of the two. For instance, assuming our EEG data was recorded from 32 electrodes at 1000 Hz and split to epochs of length 500 ms, we would have a data tensor of the shape $n \times 32 \times 500$, with n recorded epochs. Perhaps the simplest way to form feature vectors of the n epochs is by concatenating the channel and time dimensions, so that the resulting matrix will have the shape $n \times 16000$. Each row in this matrix is a spatio-temporal feature vector representation of an epoch. Furthermore, to train an LDA classifier, we need class labels. Suppose that our study had two experimental conditions: a green light displayed to the subject and a red light displayed to the subject, and that we are interested in predicting the colors of the stimuli that the subject saw based on the ERPs associated with them. We could assign labels to the epochs used for training the classifier: 0 for epochs associated with the green light, 1 for epochs associated with the red light. Of course, the epochs whose labels we are trying to predict would be unlabelled. With the epochs now in a proper shape and labelled according to the experimental condition, the LDA classifier can be defined.

Given that x_i is a feature vector representing an epoch and y_i is the corresponding class label, the binary LDA is based on the assumption that the probability distributions $p(x_i|y_i = 0)$ and $p(x_i|x_i = 1)$ are normally distributed with parameters (μ_0, Σ) and (μ_1, Σ). Note that it is assumed that the classes share the same covariance matrix Σ; by sharing the covariance matrix between classes, it is assumed that the data is linearly separable.

The classifier predicts x_i as being from class 1 if for some threshold T:

$$\ln \mathcal{L}(y_i = 1|x_i) - \ln \mathcal{L}(y_i = 0|x_i) > T, \tag{1}$$

where $\mathcal{L}$ is the likelihood function and ln is the natural logarithm. In other words, the higher the log-likelihood of $p(x_i|y_i = 1)$ compared to the log-likelihood of $p(x_i|y_i = 0)$, the more likely it is that x_i's predicted class will be 1.

By Bayes' theorem and the likelihood function of multivariate normal distributions, we get:

$$\begin{aligned} \ln \mathcal{L}(y_i = 0|x_i) &= \ln(p(x_i|y_i)p(y_i)) \\ &= -\frac{1}{2}\left((x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0) + \ln|\Sigma| + k\ln(2\pi)\right) + p(y_i), \end{aligned} \tag{2}$$

where k is the dimensionality (number of features) of x_i. Subtracting the two log-likelihoods gives us:

$$\begin{aligned} &\ln \mathcal{L}(y_i = 1|x_i) - \ln \mathcal{L}(y_i = 0|x_i) \\ &= (x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0) - (x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1). \end{aligned} \tag{3}$$

Since Σ is a covariance matrix and thus symmetric, we have that $x_i^T \Sigma^{-1} \mu_j = \mu_j^T \Sigma^{-1} x_i$, so we can write equation 1 as a dot product:

$$w \cdot x_i > c \tag{4}$$

for a weight vector w and a constant c, where

$$\begin{aligned} w &= \Sigma^{-1}(\mu_1 - \mu_0) \\ c &= \frac{1}{2}\left(T - \mu_0^T \Sigma^{-1} \mu_0 + \mu_1^T \Sigma \mu_1\right) \end{aligned} \tag{5}$$

Thus, the predicted class y_i depends on which side of the decision boundary, the hyperplane perpendicular to w, x_i lies. c defines the location of the hyperplane in the feature space. For clarity, figure 3 (left) shows a visualization of LDA performed on a toy dataset generated from two 2-dimensional Gaussian distributions $\mathcal{N}(\mu_0, \Sigma)$ and $\mathcal{N}(\mu_1, \Sigma)$ having $\mu_0 = [1.2, 0]^T$, $\mu_1 = [-1.2, 0]^T$, and a shared covariance matrix $\Sigma = [[1.5, -0.7], [-0.7, 0.8]]$.

When training the classifier, the model parameters μ_0, μ_1 and Σ are unknown and have to be estimated from the data

$$\begin{aligned} \hat{\mu}_j &= \frac{1}{m_j} \sum_{y_i = j} x_i, \\ \hat{\Sigma} &= \frac{1}{2}\left(\hat{\Sigma}_0 + \hat{\Sigma}_1\right), \end{aligned} \tag{6}$$

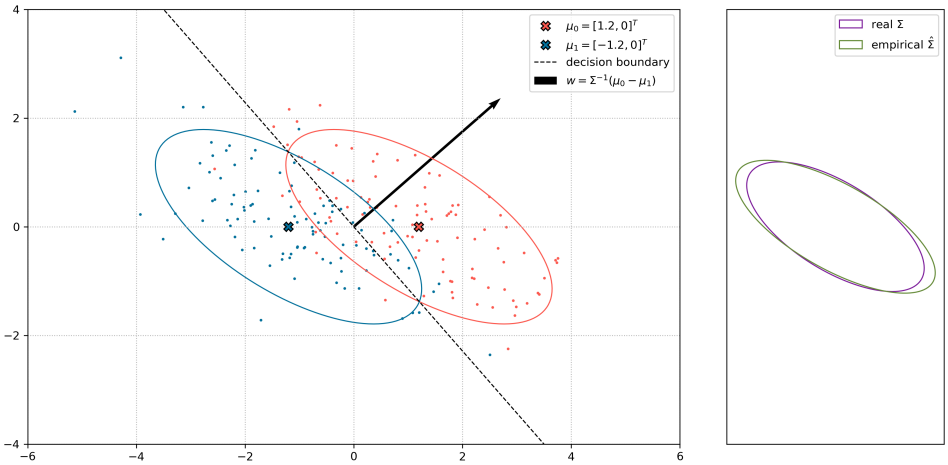


Figure 3: Left: A visualization of two 2-dimensional Gaussian distributions $\mathcal{N}(\mu_0, \Sigma)$ and $\mathcal{N}(\mu_1, \Sigma)$, with the weight vector w and decision boundary as computed by LDA. $n = 100$ for both distributions. Right: the toy dataset's real covariance matrix Σ versus its empirical covariance matrix $\hat{\Sigma}$ estimated from the data.

where m_j is the number of data points in class j, and $\hat{\Sigma}_j$ is an empirical covariance matrix computed using:

$$\hat{\Sigma}_j = \frac{1}{m_j - 1} \sum_{y_i = j} (x_i - \mu_j)(x_i - \mu_j)^T. \tag{7}$$

High dimensionality of data compared to number of samples, sometimes called the curse of dimensionality, can cause poor performance of the LDA. This is because high features per sample ratio leads to the large eigenvalues of the covariance matrix being estimated too large and small eigenvalues too small, which can be thought graphically as the excess "stretching" of the estimated distributions. Figure 3 (right) illustrates this inaccuracy for the toy dataset. To improve the classifier performance a common method called shrinkage [23] is often utilized. In this method the empirical covariance matrix is replaced with a biased one that is forced to be more spherical. Sphericalness of the covariance matrix can be thought of as the amount of regularization applied to the LDA. Indeed, another name for LDA with shrinkage is regularized LDA. The amount of regularization is controlled by a tuning parameter, for which an analytical solution is provided by the Ledoit-Wolf -lemma [45].

Because the covariance matrix is shared between the two classes, it captures the variation in the data not caused by different experimental conditions. Assuming that noise is invariant of the experimental condition, which in our case means that the same noise sources are present regardless of

the color of the displayed light, the covariance matrix should capture noise leading to improved classification performance [9]. Of course, this assumption does not always hold. If, for example, the subject blinks every time she sees a red light and never when a green light is displayed, the noise caused by the blinks is correlated with the red light condition and the covariance matrix fails to capture the noise.

3 Informativeness of words in natural language

Language is a system of symbols that enables us to communicate the concrete and customary as well as the abstract and astonishing. When we read text or listen to speech, our brains interpret the meaning embedded in language in real time, and produce a mental representation of the information that is being conveyed. The smallest element of language that can carry meaning in isolation is a word [35]. Words vary in their informativeness. For instance, the word 'Gandhi' likely induces an image of a bald man with round glasses in the mind of the reader, while the word 'since' arguably does not. Informativeness, as we use the term here, is a measure on how well a word describes some topic. Continuing our example, 'since' occurs in a variety of documents, while the document in which 'Gandhi' appears most likely has something to do with the Indian political activist. These words also have statistical differences: the frequencies of the word 'since' across documents are likely quite uniform, while the occurring frequency of the word 'Gandhi' definitely varies from document to document. Furthermore, the practical usefulness of these statistics is demonstrated by their wide exploitation in information retrieval applications. Since the informativeness of words differ intuitively as well as statistically and their statistical features have proven to be valuable in real-world applications, the question we ask in this work is: Is this statistical difference reflected in neural activity associated with reading text?

3.1 Language models

Language Models (LMs) provide a statistical method for modelling language. They are not concerned with the syntax or semantics of language, and can be computed from any body of text without knowledge of the text's structure or its meaning. Language models assign a probability for the occurrence of a sequence of words $P(w_1, w_2, ..., w_n)$. The unigram language model assumes that each word occurs independently of each other, formally $P(w_1, w_2, ..., w_n) = P(w_1)P(w_2)...P(w_n)$. This probability can be estimated from text using only word frequencies. Despite its seeming simplicity, the unigram LM performs well empirically and is a popular choice in information retrieval applications [80]. This work is concerned with unigram language models, and the terms language model and unigram language model are used interchangeably in the remainder of this work.

Language models were introduced to the information retrieval audience in 1998 by Ponte and Croft [63], who showed that language models can be used to compute a query likelihood model, which assigns the generative probability $P(Q|d)$ for a query $Q = (w_1, w_2, ..., w_3)$ and a document d. The query likelihood model, they demonstrated, could be used to estimate the relevance of a document with relation to a query, and that the retrieved documents could be ranked based on the estimated relevance.

3.2 Model for estimating word informativeness

Here, we present a statistical model for estimating word informativeness from a collection of documents part of a larger corpus of text. The informativeness of a word is estimated with the entropy of the distribution of word's generative probabilities over documents in the collection. To compute the generative probabilities, we build our model on the query likelihood model and language models.

We start by introducing a simple derivative of the query likelihood model, the word likelihood model. The word likelihood model assigns a probability $P(w|M_d)$ for a word w and a document model M_d. A document model is a bag-of-words representation of a document, in which the order of the words is disregarded, and only the number of occurrences of each word (word frequency) is preserved. The probability of a word occurring in a document can be estimated with the word frequencies. More formally,

$$\begin{aligned} P(w|M_d) &= \frac{f_{w,d}}{f_d} \\ \sum_{w \in M_d} P(w|M_d) &= 1, \end{aligned} \tag{8}$$

where $f_{w,d}$ stands for word frequency for word w in document d and wf_d is the total amount of words in d. Essentially, this is the probability of a document generating a single word.

However, problems arise if a word does not appear in a document, as such a word will have zero probability. Zero probabilities are problematic, since not all words are equally unlikely even if they do not appear in a document. Consider, for example, a document about cats. Assuming that neither of the words 'mouse' or 'Bangladesh' appear in said document, they have zero probabilities. However, intuitively it seems that 'mouse' is more likely related to such a document. To counter the problem of zero probabilities, several smoothing techniques have been developed, of which the well-known linear interpolation model is introduced here [67]. The linear interpolation model combines the document-specific models with a corpus model. The corpus model M_{corpus} is constructed similarly as the document models, and it consists of the word frequencies in the whole corpus. This way, the corpus model gives the probability for a word appearing in the corpus, regardless of the document. The document-specific models and the corpus model are combined to form a smoothed likelihood P_s:

$$P_s(w|M_d) = (1 - \lambda)P(w|M_d) + \lambda P(w|M_{corpus}), \tag{9}$$

where $0 < \lambda < 1$ is the smoothing parameter; high values of λ mean more smoothing. The smoothing parameter may be chosen freely, but generally speaking longer queries require more smoothing and shorter less

[81]. Since in our case the "query" consists of only one word, low λ values are recommended. In essence, smoothing transfers some of the probability mass of the most probable words to the less probable ones, smoothing the probability distribution of the words towards their probability distribution given the corpus.

Finally, since we are interested in the distribution of documents given a word, we calculate $P(d|w)$. By utilizing Bayes rule this becomes:

$$P(d|w) \propto P(w|d)P(d), \tag{10}$$

where $P(t)$ can be ignored, since it is the same for all d. If it is assumed that the documents have an uniform prior probability, the equation can be simplified further:

$$P(d|w) \propto P(w|d) \tag{11}$$

Due to this, $P(w|d)$ can be used to compute the probability of a word "generating" a document.

To illustrate how the smoothed word likelihood model works, consider an example where document models are constructed from all of Wikipedia's articles. The corpus model M_{corpus} contains all of the documents (articles) in Wikipedia. Let smoothing parameter $\lambda = 0.1$. Table 1 shows examples of word likelihoods for two documents in the corpus, India and Cat. We see that in the India document, the word 'cat' had a zero probability prior to smoothing, and non-zero probability after smoothing. Furthermore, the word 'cat' is much more likely to appear in the cat document than it is in the India document. On the other hand, the word 'republic' is more likely to appear in the India document than in the cat document. Moreover, we see that words that are common in the corpus, 'the' and 'from', have likelihoods of similar magnitude in both of the documents, whereas the likelihoods of the domain-specific words 'cat' and 'republic' vary highly with relation to the document. Intuitively, such words can be thought to discriminate the documents from each other.

word (w)	$P(w\|M_{corpus})$	$P(w\|M_{india})$	$P_s(w\|M_{india})$	$P_s(w\|M_{cat})$
the	0.069	0.056	0.0573	0.0321
cat	4×10^{-05}	0	4×10^{-06}	0.0252
from	0.005	0.004	0.0041	0.0032
small	0.0004	0.0003	0.0003	0.0018
republic	0.0002	0.0010	0.0009	7.4×10^{-05}

Table 1: Examples of word likelihoods of a smoothed query likelihood model constructed from Wikipedia articles. From left to right: word, likelihood of word in the document about India, smoothed likelihood of word in document about India, smoothed likelihood of word in the document about cats.

We can now compute the probability distributions for the words. Figure 4 shows the probability distributions over 30 documents in the corpus for the words 'the', 'small' and 'cat'. The probabilities have been normalized so that they sum to unity.

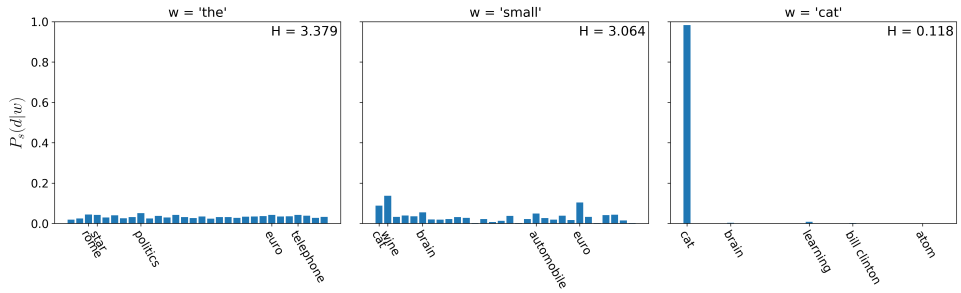


Figure 4: Probability distributions over 30 documents for the words 'the', 'small' and 'cat'. Entropies (H) of the distributions are shown in the upper right corner of each plot. The five most probable documents are labelled.

As can be seen from the figure, distributions look very different for different words. Distribution for the word 'the' resembles a uniform distribution, whereas the distribution for 'cat' has a clear mode of document 'cat'. This is reflected in the entropies of the distributions. In order to measure the informativeness of an individual word w_i in the context of the document collection, we calculate the Shannon entropy of the distribution

$$H(d|w_i) = -\sum_d P(d|w_i) \log_2 P(d|w_i)$$

The entropy $H(d|w)$ gives us an estimate of word informativeness: low entropy signifies high informativeness (certainty of a document given a word), and high entropy signifies low informativeness (uncertainty of a document given a word).

4 Neurophysiological experiment on word informativeness

This chapter presents a neurophysiological experiment conducted to study the relationship between human brain signals and the informativeness of words. The aim of this experiment is to either confirm or reject the hypotheses that brain activity associated with reading natural language text is affected by the informativeness of the read words (H1). For this purpose, the EEG data of 15 participants reading Wikipedia articles is analysed. Word informativeness is computed for the words displayed to the participants, and the effect of word informativeness to ERPs associated with the words is visualized and studied with statistical methods.

4.1 Methods

This chapter is structured as follows. The first section describes the experiment in which the EEG data was measured, followed by a section detailing the preprocessing steps taken to improve the signal-to-noise ratio of the EEG recording. In the section following these the informativeness of the words used in the EEG measuring experiment is estimated with the model defined in chapter 3.2. The final section describes the statistical models used for testing the significance of the findings and confirming/rejecting the hypotheses H1.

4.1.1 EEG data measurement experiment

The neurophysiological experiment was performed with the raw EEG data from a previously published study [16]. The aim of the study was to build an information retrieval system that utilizes EEG to recommend documents relevant to participants' interests. A brief description of the experimental procedure of the study follows. The full disclosure along with illustrations of the experimental setup can be found in appendix A.

General setup In the study 15 participants read documents from a pool of 30 documents while their EEG was measured. These 30 documents were articles in the English Wikipedia. The titles of these articles are displayed in Table 2. All participants completed eight reading tasks, each of which consisted of reading two documents. Before the start of each reading task, the participants chose freely which of the two documents was relevant to them, and their answers were stored alongside their EEG data. The participants were instructed to keep in mind the topic of the relevant document during the whole reading task.

Reading task A reading task consisted of six trials, and during each trial two sentences, one from the irrelevant and one from the relevant topic were shown to the participant. Thus, the six first sentences of the documents

were shown to the participant. The sentences were displayed on a computer screen one word at a time with 699 ms between words. The participants were instructed to simply read what was displayed to them, and asked to not engage in any additional tasks or mental imagery. The screen was masked by a black rectangle with a grid-like pattern and an opening for the word. This was used to control the degree to which word length affected light reaching the eyes (i.e. to make sure longer words did not produce more black pixels on the screen). Furthermore, punctuation marks were omitted, and the sentences were separated with a mask of word-like sequences consisting of 4 to 9 numbers (`1111`) or other non-alphabetic characters (`&&&&&&&`) in the opening where the words appeared. At the end of each trial and reading task, the participants answered various questionnaires that were relevant to the original study, but which were not exploited in the present study.

Atom	Cat	Machine learning	Politics	Star
Automobile	Communism	Michael Jackson	Rome	Telephone
Bank	Euro	Money	Savanna	Time
Bicycle	Association football	Ocean	Schizophrenia	Volcano
Bill Clinton	India	Painting	School	Wife
Brain	Learning	Plato	Society	Wine

Table 2: Titles of the 30 documents used in the EEG experiment.

4.1.2 EEG preprocessing

The EEG data consisted of the EEG signal recorded from 32 electrodes placed on the participant's scalp (denoted as C), along with markers indicating the precise onset times of the stimuli. The labels and placement of the electrodes is shown in Figure 2 (left). The data sampling rate was 2000 Hz. To remove low-frequency signal fluctuations and high-frequency line noise the data were band-pass filtered at the frequency range of 0.25 - 35 Hz with a Firwin1 filter. After this, the data were split to epochs spanning -200 - 1000 ms relative to the onset of each stimulus. This resulted to a data tensor $X^{|N|\times|C|\times t}$ with $N = [0, ..., n-1]$ epochs, $|C| = 32$ EEG channels, and $t = 2400$ datapoints, as measured with 2000 Hz for 1200 ms epochs. Another dataset, denoted X', was created from X for computing per-participant thresholds that were used to identify bad channels and epochs in X. X was used in all of the experiments in this study, while X' was discarded once data preprocessing was complete.

The epochs in X' were absolute baseline corrected using the time range of the whole epoch, -200 - 1000 ms. Next, the following subset of channels was picked for artefact detection:

$$C' = \{F3, Fz, F4, FC1, FC2, C3, Cz, C4, CP1, CP2, P3, Pz, P4\}$$

This subset of channels was chosen, because these channels reside near the

top of the head, where less noise is present. Bad channels and epochs were identified using a voltage threshold. This threshold was calculated for each participant separately, because individual factors affect the voltages measured by the electrodes. Using the aforementioned channels, a maximum absolute voltage v_i was calculated for each epoch $i \in N$ for the time interval -200 to 700 ms. Formally,

$$v_i = max(|x_i'^{|C'|\times t'}|), \tag{12}$$

where the matrix $x_i'^{|C'|\times t'}$ represents epoch i's data points for channels C' and times $t' = 1800$ (2000 Hz for 900 ms epochs) from dataset X'. This time range was chosen since the epochs overlapped in the range 700 - 1000 ms after stimulus onset, and invalidating two epochs due to one artefact was to be avoided.

The 80th percentile of the absolute max voltages v_i was assigned as the voltage threshold v_{thres}. The threshold values ranged from 25 to 67 μV between subjects (for full disclosure, see appendix B). Epochs with an absolute maximum voltage over the threshold were marked as bad. Formally,

$$N_{bad} = \{i \in N \mid v_i > v_{thres}\} \tag{13}$$

In other words, 20% of the epochs with the highest absolute maximum values were deemed bad. Note that the epochs were not removed from the data at this point, because they were needed to find bad channels.

To find bad channels caused by for example a loosely fitting cap or a malfunctioning electrode, the absolute maximum voltage was computed separately for each channel and epoch in X'. Each epoch with an absolute maximum voltage over v_{thres} as well as epochs with a voltage variance less than 0.5 μV were considered bad. Furthermore, channels with bad epoch rate of over 20% of all epochs were marked as bad:

$$C_{bad} = \left\{ c \in C \mid \sum_{i=0}^{n-1} I(max(|x_{ic}'^{t'}|) > v_{thres} \text{ or } var(x_{ic}'^{t'}) < 0.5) > 0.2n \right\},$$

where $x_{ic}'^{t'}$ is the vector of data points for epoch i and channel c on the time range t'. I is the indicator function, and var gives the variance of its argument vector.

Finally, the following modifications were made to the final data set X: the bad epochs N_{bad} were dropped, and the bad channels C_{bad} were interpolated using spherical splines [60], and each epoch was absolute baseline corrected using the pre-stimulus period -200 - 0 ms. After the preprocessing the average number of epochs per participant was 1550. The per-participant interpolated channels and numbers of dropped epochs can be found in appendix B.

4.1.3 Estimation of word informativeness

Section 3.2 described a model with which informativeness of words could be estimated from a collection of documents in a corpus. Word informativenesses used in the rest of this thesis were estimated from the English Wikipedia using the model. Document models of 30 articles were generated as well as a corpus model consisting of all of Wikipedia's articles. These documents coincided with the ones used in the EEG measuring experiment. Prior to constructing these models punctuation marks were removed from the text and the words were stemmed using the Porter stemming algorithm [64]. The Porter stemmer removes the suffixes of words, attempting to map words with similar meanings to one word. For example, the following words:

```
connect, connected, connecting, connection, connections
```

all map to the stem `connect`. Stemming and removal of punctuation marks (tokenization) are both common procedures in natural language text processing.

A smoothed word likelihood model was constructed using the aforementioned models. Using these models, informativeness was computed for each of the stemmed words occurring in the 30 documents. Since λ values around 0.1 have been shown to produce the best results on short queries when using query likelihood models in information retrieval [81] and the "query" length in the case of a word likelihood model is effectively one, $\lambda = 0.1$ was chosen as the smoothing parameter. Words with an estimated informativeness in the 25th percentile were labelled as informative words (label 1), and words with estimated informativeness greater than the 25th percentile uninformative words (label 0). A histogram of the occurrences of informativenesses can be seen in Figure 5. The theoretical maximum entropy for the set of 30 documents is $\log_2(30) \approx 3.401$, and it is reached when $P(d|w)$ is uniform.

In order to check the validity of the model, the words were labelled by three human assessors to be either relevant (label 1) or irrelevant (label 0) to the document from which they were from. The annotating setting was single-blind, so the assessors were unaware of the words' informativenesses at the time of labelling. Approximately 25% of the words were labelled relevant, so in order to match the class sizes of the human annotated labels and informativeness labels, the 25th percentile cutoff for informative/uninformative classes was chosen. The human annotated relevance and the informativeness labels were found to be in substantial agreement (Cohen's κ 0.612).

4.1.4 Statistical significance testing

Independence of observations is an assumption in the popular Analysis of Variance (ANOVA) models, and breaking this assumption may lead to overconfidence of the test (high Type I error rate). The experimental setup in the present study provides many factors which make the observations

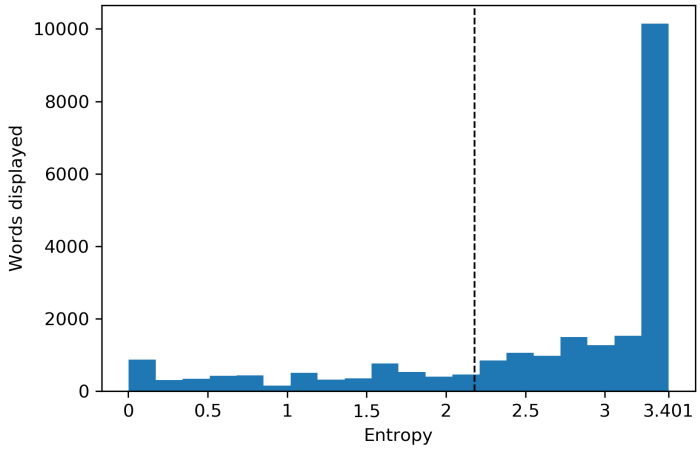


Figure 5: Occurences of informativeness values for all words presented to the participants, with the dashed line marking the 25th percentile of the informativeness values.

non-independent. The usage of natural language as stimuli introduces non-independencies such as: the words occur in a context, and the context may affect the interpretation of a word; word lengths may be correlated with ERP measurements; and, the participants freely chose the documents they read, which causes different stimuli being presented to participants. Additionally, non-independencies are introduced by common factors in neuroscientific studies, such as inter-subject variance in measurements. In order to avoid excessive Type I errors, Linear Mixed Models (LMMs) were used for statistical testing, as they allow for partial relaxation of the independence assumption by random effects.

Linear Mixed Models LMMs are gaining popularity in natural sciences, including neuroscience [10]. LMMs are an extension of linear models, i.e. linear regression. Like linear regression, LMMs have fixed effects and a random noise variable that explain the dependent variable - measurement in our case. Additionally, LMMs have random effects, which are like fixed effects, but but instead of using the fixed effects themselves, a normal distribution is estimated for each of them.

For example, consider a simple neurophysiological experiment with one experimental condition, say, a binary stimulus type, and one participant. The experiment has n trials, meaning that n stimuli were presented. To study the effects that the stimulus type has on a certain ERP-component, we could fit a linear regression model with the stimulus type as a fixed effect

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{14}$$

where y_i is the measurement (ERP-component); $x_i \in \{0, 1\}$ is the stimulus type of trial $0 \leq i < n$; β_0 is the intercept; β_1 is the slope of the fixed effect; and $e \sim \mathcal{N}(0, \sigma^2)$ is the error term with a variance of σ^2. By fitting this model to the data, β_1 would reveal us the possible correlation that the experimental condition has on the measurements. Suppose now, that our experiment had multiple participants, which introduces a non-independency to the measurements. Modelling this data with the previous model would lead to large error terms and a poor model fit. To improve the model, we introduce a random variable $S_j \sim \mathcal{N}(0, \tau^2)$ for each subject j, which models variance caused by inter-subject differences. Thus, we have

$$y_{ij} = \beta_0 + S_j + \beta_1 x_i + e_{ij} \tag{15}$$

The random variable S_j allows the predictions for each subject to be adjusted either upward or downward. Now, our model contains fixed effects β_0 and β_1, as well as a random effect S_j. Thus, it is called a linear mixed model. Essentially, adding random effects adds structure to the error term by estimating multiple distributions that cause the variance in the measurements. Intuitively, the fact that subject is represented as a random variable in the model reflects our uncertainty of how inter-subject differences affect the measurements.

To improve our model, more experimental conditions could be taken into consideration by adding fixed effects, and further non-independencies of the measurements could be solved by adding random effects.

Model used for significance testing Linear mixed models were used to analyse the relation between word informativeness and ERP components. The mean voltage in the Pz channel was computed for each ERP component (components and their time windows specified at the beginning of the results section), and LMMs were fit for the data corresponding to each of the components.

To avoid common pitfalls resulting to Type I errors, the LMM models were designed using the "keep it maximal"-principle presented by Barr et al. [4]. Fixed effects in these models were word informativeness (continuous) and topic relevance (as specified by the participant in the EEG measurement experiment, binary). Random effects included intercepts for subjects, word lengths and topic of the document from which the word was from; also a by-subject random slope for the effect of word informativeness was included. Visual inspection of residual plots did not reveal obvious deviations from

normality or homoskedasticity. The models were evaluated with likelihood ratio tests of the alternative hypotheses model with word informativeness as a fixed effect, and a null hypothesis model, which excluded the informativeness effect, but was otherwise the same. As an example, the following is the specification of the alternative and null LMMs (respectively) for the N400 component, in R-typical notation:

```
N400 ~ t_info + topic_rel + (1|subject) + (0+t_info|subject) + (1|wordlen) + (1|topic)
N400 ~ topic_rel + (1|subject) + (0+t_info|subject) + (1|wordlen) + (1|topic),
```

where `t_info` stands for informativeness and `topic_rel` for topic relevance.

A significant inverse correlation between word informativeness and word length (Spearman's ρ -0.74, $p < 0.0001$) was observed. Word length was included as a random effect in the model in order to mitigate its effect on the results.

4.2 Results

Combining the informativeness measures of the words and the EEG data, the effect of informativeness of words to event-related potentials was studied (H1). The following ERP components were investigated: P200 [100, 250] ms, P300 [250, 350] ms, N400 [350, 500] ms, and P600 [500, 800] ms. These time intervals were chosen based on visual inspection of the ERPs and literature [30, 68, 38, 58]. The uninformative/informative word classes were used for visualizations, and the statistical significance testing was conducted with continuous word informativeness values.

The grand-average topographic scalp plots in Figure 6 show that there is a difference between brain activity associated with words in the informative/uninformative classes. The plot depicts the scalp topographies for the two classes at times 175 ms (P200), 300 ms (P300), 425 ms (N400) and 650 ms (P600), with averages of time windows of 100ms (so, for example, the plot at 175 ms is an average of data points in the time interval 125 - 225 ms). The topographies are averaged over all participants and stimuli words. Also depicted are the differences between the informative/uninformative classes' topographies for the aforementioned times. The greatest difference between the classes is visible near the Pz electrode and for ERP components P300 and N400.

The largest divergence between the two classes was found in the Pz electrode, and it was therefore chosen for further inspection. Figure 7 shows the grand-average ERPs over all participants and stimuli words at the Pz electrode (ERP plots for all channels can be found in Figure C1). The following listing provides interpretations of Figure 7. The ERP-components which were significantly affected by word informativeness are marked with a star (*).

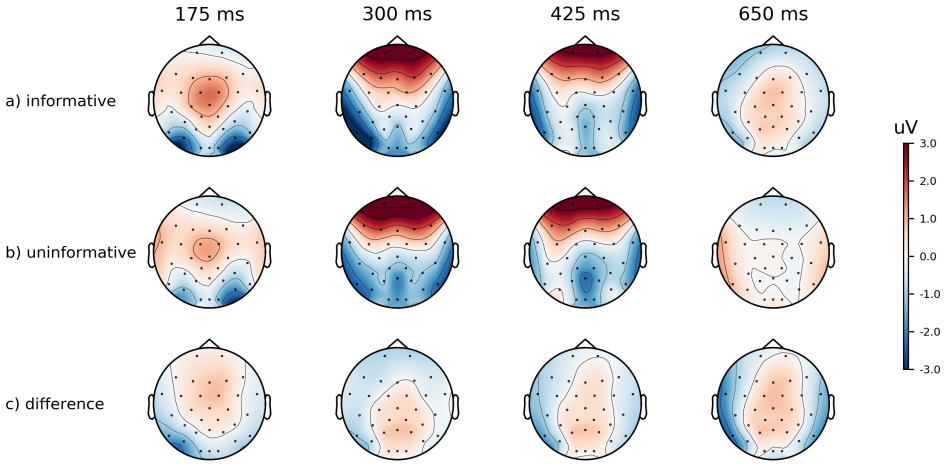


Figure 6: Grand average-based topographic scalp plots of ERPs from [125-225] ms, [250-350] ms, [375, 475] ms, and [600, 700] ms after word onset: a) ERPs associated with informative words, b) ERPs associated with uninformative words, c) ERPs associated with informative words minus ERPs associated with uninformative words. Each contour line in the images represents a $1\mu V$ difference in voltage.

P200* The voltages for words in the informative/uninformative classes differ briefly at the P200 component. The P200 component is affected by the physical features of the stimuli [57], so the difference in the amplitudes for the two informativeness classes may be explained by the effects that word length has on the visual characteristics of the stimuli.

P300* The voltages for the two word classes start to diverge at the P300 component, and this difference remains until the next stimulus. The peak of the P300 component is masked by the simultaneously occuring negative N400 peak. The large variance may be explained by individual differences among participants and differences in the read documents, since individual factors as well as task difficulty affect the latency and amplitude of the P300-component.

N400 While there is a difference in the voltages of the informativeness classes, there is also the most variance compared to other components, as depicted by the shaded red and blue areas. Note that the difference between the two classes does not grow after the P300 component. Thus, the N400 component is not affected by the word informativeness.

P600* The differences in the ERP that began at roughly 300 ms post-stimuli is most clearly visible in the P600-component. While the difference

remains roughly the same during 300-800 ms post-stimuli, the variances are the lowest at the P600-component. This lower variance may be explained with the P600-as-P300 theory [65]: while the latency of the P300-component varies among participants/tasks, its peak lasts long leading to low variance and high difference in voltages at the P600-component.

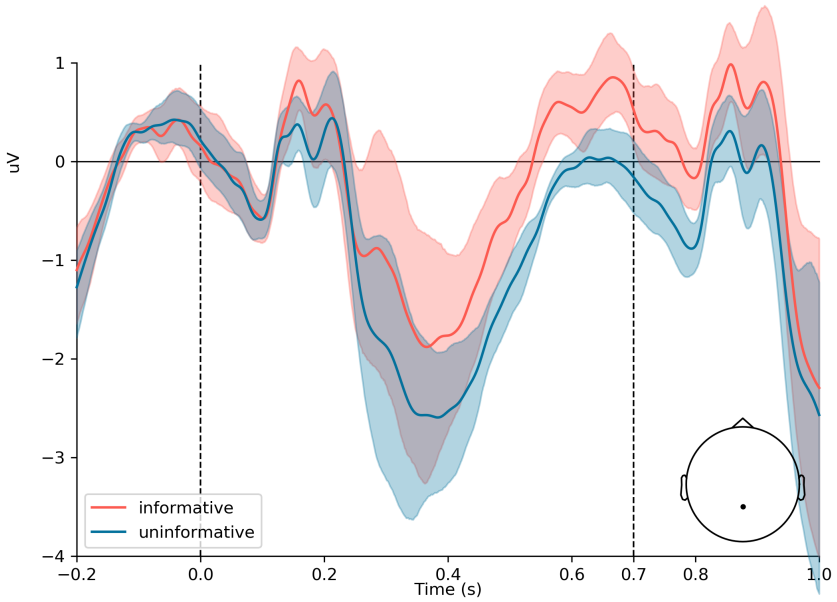


Figure 7: Grand average event-related potential at the Pz channel for informative words (red curve) and uninformative words (blue curve). The shaded areas represent a 95% confidence interval computed using a percentile bootstrap method. Word onsets are marked with dashed lines.

Table 3 shows the results of the likelihood ratio tests of LMMs for each of the ERP-components. Additionally, the topic relevance which was defined by participants in the EEG measurement experiment was tested for significant effect, but none was found in any of the components.

In conclusion, the visualizations in this chapter show a visible difference between ERPs associated with informative and uninformative word classes, and the statistical significance tests on the ERP-components show that word informativeness has a significant effect on the ERP-components P200, P300, and P600, confirming the hypotheses that informativeness of read words affects brain activity (H1).

Component	p-value
P200	**0.0097**
P300	**0.0098**
N400	0.0265
P600	**<0.00001**

Table 3: Results of likelihood ratio tests between the alternative and null LMM models. Significant ($p < 0.05$) values after Bonferroni-correction ($m = 4$) are highlighted.

5 Predicting word informativeness from ERPs

To investigate whether the informativeness of words can be predicted from the brain activity associated with reading the words (H2), a single-trial classification experiment was conducted. The experiment used the same preprocessed EEG dataset as the neurophysiological experiment on word informativeness (sections 4.1.1 and 4.1.2). Single-trial classifiers were trained for each participant separately. The classifiers were trained with the epoched EEG data and binary labels indicating whether the ERPs of each epoch were associated with an informative or an uninformative word. The uninformative/informative classes were the same as defined in the neurophysiological experiment (section 4.1.3). To test the hypothesis H2, the classifiers were evaluated against classifiers trained with permuted labels.

5.1 Methods

The first section in this chapter describes the classifier and data representation used to train it, followed by a section detailing the steps taken to ascertain whether the classifier could predict the word informativeness.

5.1.1 Classifier details and feature engineering

Linear Discriminant Analysis (LDA) was utilized to learn linear classifiers to separate ERPs associated with words in the informative and uninformative classes. LDA has been previously used successfully in single-trial ERP classification [52, 46]. Binary LDA classifiers were trained for each participant separately due to the individual differences in EEG measures. Since we wanted the classifier to utilize both the spatial attributes (channels) as well as the temporal attributes (time w.r.t. stimulus onset) of the data, some feature engineering had to be carried out. The tensor $X^{m \times |C| \times t}$ represents the preprocessed EEG recording for each participant, with $m = |N| - |N_{bad}|$ cleaned epochs, $|C|$ channels and t time points. To reduce the dimensionality of the data, the time points were divided to $t' = 8$ equidistant windows between 150ms and 950ms, and the average voltage of each of these windows was computed, resulting in a $X^{m \times |C| \times t'}$ tensor. This led to time windows spanning 100ms. The time intervals were chosen based on the ERP analysis of section 4.2. Furthermore, the channels and time windows were concatenated

together, resulting in a $X^{m \times |C| \cdot t'}$ spatio-temporal feature matrix. This feature engineering procedure follows standards for single-trial ERP classification [9]. Since the data is of a relatively high dimensionality ($32 \cdot 8 = 256$) compared to the number of data points (approximately 1400 per training set), LDA with shrinkage was employed. The tuning parameter for shrinkage was chosen with the Ledoit-Wolf -lemma [45].

To be able to evaluate the classifiers, the epochs of each participant were split to eight blocks $B = \{b_0, ..., b_7\}$ coinciding with the eight reading tasks in the EEG measurement experiment. Consequently, each block consisted of the epochs for two documents. A classifier was trained for each block b_i so that each of these classifiers used seven of the other available blocks as a training set $X_{\{B \setminus b_i\}}^{(m-m_i) \times |C| \cdot t'}$, and were evaluated on the test set $X_{b_i}^{m_i \times |C| \cdot t'}$.

These classifiers were trained with the informativeness labels (informative: 1, uninformative: 0). The split at the 25th percentile resulted in imbalanced classes; however, LDA has been shown to be quite robust against class imbalances ([77], but see [76]).

5.1.2 Classifier performance evaluation

The performance of the classifier was measured with the Area under the ROC curve (AUC). This measure was chosen because AUC combines the true positive and false positive rate, and thus gives better estimates when the classes are imbalanced. In the case of imbalanced classes, the classifier will tend to predict the dominant class (in this case the uninformative class), which causes accuracy to give overconfident estimates of performance.

The classifier performance was evaluated with permutation tests. The classifier was trained with permuted class labels to reveal if the classifier had learnt any real class structure in the data. With a sufficiently high number of permutations this produces permutation-based p-values [56]. The null hypothesis is that the class labels and brain activity are independent of each other. A small p-value indicates that the classifier is able to find some meaningful structure of the brain activity that correlates with the class labels (informative/uninformative). We ran $k = 1000$ permutations for each subject, so k classifiers with randomly permuted labels were trained for each subject, and their AUCs were compared to the AUC of the actual classifier to produce the p-values. To obtain the AUCs for each subject, we calculated the mean of the AUCs of the per-block classifiers.

5.2 Results

The average classifier AUC score over all participants was 0.643. The classifiers of all subjects performed significantly better than random-permutation classifiers (AUC = 0.5, $p < 0.01$). In other words, the classifiers were able to predict the word informativeness from the ERPs with a performance better than a random baseline, confirming the hypotheses H2. The AUC scores of the classifiers can be seen in Figure 8. Since the participants of the EEG

measuring experiment were instructed to not engage in any mental imagery or tasks besides reading the text displayed to them, conscious control of brain activity was not required to make prediction of word informativeness possible. This means that the predictions of the classifier could possibly be utilized in a passive BCI application.

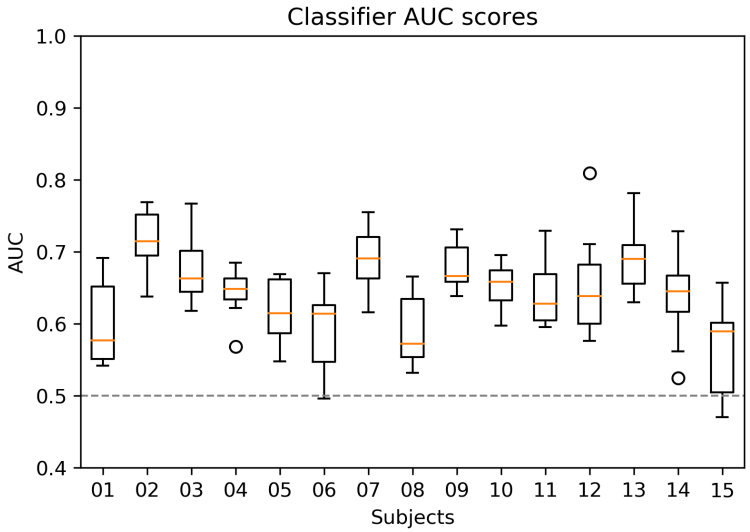


Figure 8: The classifiers' AUC for each subject and block. The dashed line marks the performance of a random-permutation classifier.

Illustrating classification results, Table 4 shows top/bottom 5 predicted words in the informative/uninformative classes for three randomly selected documents. The prediction confidence for words to belong to the informative class was computed over all participants and reading tasks. The first two columns show the five words which had the highest predicted word informativeness alongside with the five words with the highest actual word informativeness (ground truth) in the selected documents. In these columns, the words are sorted in a descending order based on classifier confidence or informativeness, respectively. The remaining two columns show the five words which had the lowest predicted word informativeness and lowest actual word informativeness. These words are sorted in an ascending order based on classifier confidence or informativeness, respectively. For example, the word 'shizophrenia' had the highest classifier confidence of all the words in the document, and the word had the 3. highest informativeness of the words in the document.

There is a clear difference between the words in the informative/uninformative predicted classes. The words in the high predicted column tend to be related to the topic of the document (e.g. 'asia', 'civilisation', 'mahatma' for document india), while exceptions exists (e.g. 'additional' in schizophrenia). On the other hand, words in the uninformative predicted class tend

to be short functional words, which are intuitively not very informative. To summarize, the classifiers seem to have captured the differences in ERPs of informative and uninformative words quite well. The full list of top 5 predicted/ground truth words for all documents can be found in appendix D in table D1.

topic	informative predicted	informative true	uninformative predicted	uninformative true
cat	mammal	felids	as	they
	housecat	housecat	and	as
	indoor	felines	no	is
	despite	cats	in	in
	flexible	cat	too	and
india	asia	indus	in	is
	civilisation	multilingual	vast	in
	independent	gandhi	to	and
	subcontinent	mahatma	led	to
	mahatma	civilisation	of	the
schizophrenia	schizophrenia	antipsychotic	primarily	is
	environment	hallucinations	what	and
	characterized	schizophrenia	by	to
	additional	contributory	a	the
	psychological	symptoms	of	of

Table 4: Top/bottom 5 words per topic sorted by classifier confidence (predicted) for class membership (informative/uninformative) and by actual informativeness i.e. ground truth (true).

Feature analysis To gain insight in what features of the data discriminate the informative/uninformative classes, a separability index matrix was composed. Signed-r^2 values chosen as the separability index. The signed-r^2 values (r^2_{sgn}) are the signed squares of the Point-Biserial correlations of each feature with regards to the real class label. Point-Biserial correlation coefficient r is calculated as

$$r(x_j, y) = \frac{\sqrt{n_0 n_1}}{n_0 + n_1} \frac{\text{mean}(x_j \mid y = 0) - \text{mean}(x_j \mid y = 1)}{\text{std}(x_j)}, \tag{16}$$

where j is the index of the feature, n_k is the number of samples in class k, and y is the class label. x_j and y are vectors. The signed-r^2 values are defined as $r^2_{sgn}(x_j, y) := \text{sgn}(r(x_j, y)) \cdot r(x_j, y)^2$. The r^2_{sgn} values were then organized in a matrix and an average of these matrices was computed over participants. This result is plotted as a color-coded image in Figure 9 (left). Warmer colours in the plot indicate that the feature in question correlates positively with the informative class, whereas colder colours indicate a negative correlation between the feature and high informativeness Shale colors implicate that the feature in question has little correlation with the

class label. For example, the dark-orange colour at the Pz channel at time window 4 indicates that on average the voltage at the Pz electrode at time 450 - 550 ms post-stimulus correlated positively with the class label 1, which stands for high informativeness of the word displayed.

For clarity, Figure 9 (right) shows the signed-r^2 values as topographic scalp maps. Each map corresponds to a time window in the feature space. For example, the top left image corresponds to the time window 150 - 250 ms, which is the first column in Figure 9 (left). The figures show that the activity by which the ERPs are classified occurs mainly in the central electrodes, and that the frontal (Fp1, Fp2) and the occipital (O*) electrodes contribute less to the classification task.

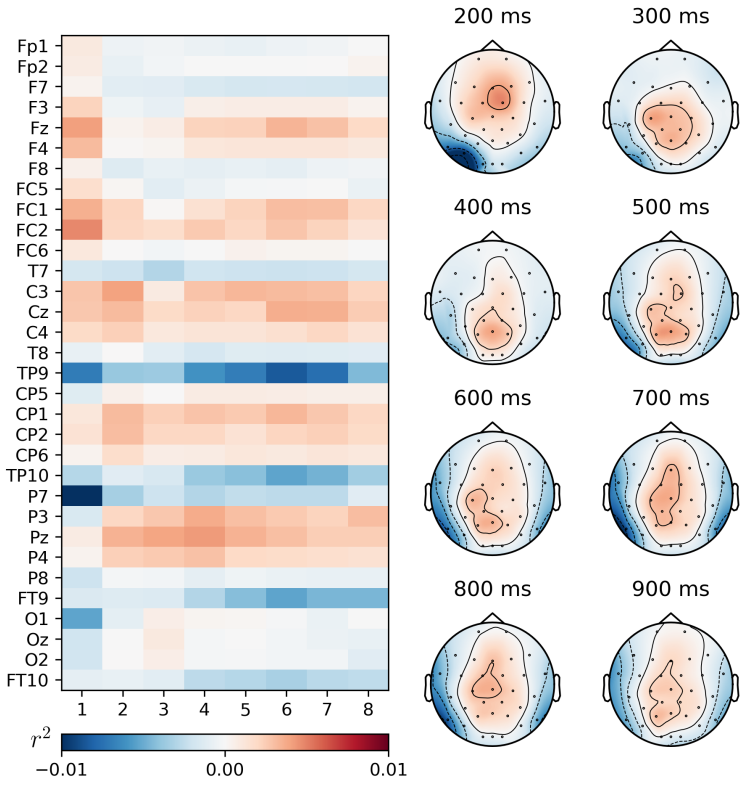


Figure 9: Average of per-participant r^2 matrices; left: r^2-values plotted as an image with channels on the vertical axis and time windows on the horizontal axis, right: r^2-values plotted as scalp topography maps, with each map corresponding to a time window.

Noise matrix Figure 10 displays the average of the covariance matrices (Σ) of each participant's classifier. Intuitively, the covariance matrix of the classifier accounts for the noise in the data; it captures the variation in the ERPs that is not captured by the difference in class means. Remembering the

definition of the weight vector of the LDA classifier: $w = \Sigma^{-1}(\mu_1 - \mu_0)$, we see that the covariance is "divided out" from the difference of the distribution means. This aligns w so that it separates the classes efficiently regardless of the skewing of the class distributions caused by noise (see Figure 3 [left]). In this context, noise is defined as any voltage fluctuation not resulting from the experimental condition (informativeness). It is worth noting that this may also include brain activity produced by other cognitive processing not related to the task at hand.

Moreover, the covariances show how channels affect each other. For example, the positivity in the Fp1 and Fp2 channels is reflected as a positivity in the frontal electrodes that reside near them and as a negativity in the other electrodes. These positive voltages are likely due to eye blinks, which cause large voltage fluctuations in the frontal electrode sites. In general, channels in close proximity to each other tend to have a positive covariance, whereas on channels more remote from each other the covariances are negative. This is explained by the fact that when average referencing is used the total scalp voltages sum to zero: positive activity in some areas mean negative activity in another.

In addition to spatial covariances, the covariance matrix displays temporal covariances. In the image, there are 32x32 smaller squares visible. Each square contains 8x8 dots depicting the covariances between time windows for a particular channel. For example, the top left square contains the covariances of each of the $t' = 8$ time windows in the Fp1 electrode. In that electrode, the time windows 2-4 (250-550ms) appear to have the highest variance. We also observe that time windows in close proximity with each other covary with each other.

Note also the diagonal line of squares in the matrix. These show the variances of each channel. The channels residing near the top of the head have less variance than those residing on the sides. Furthermore, there seems to be more variance towards the end of each time window. This can be seen with the green colour darkening towards the bottom right corner of each channel-square. It is noteworthy, however, that channels and time windows which appear noisy in the averaged covariance matrix may be noisy for only some participants and provide clean signal on others.

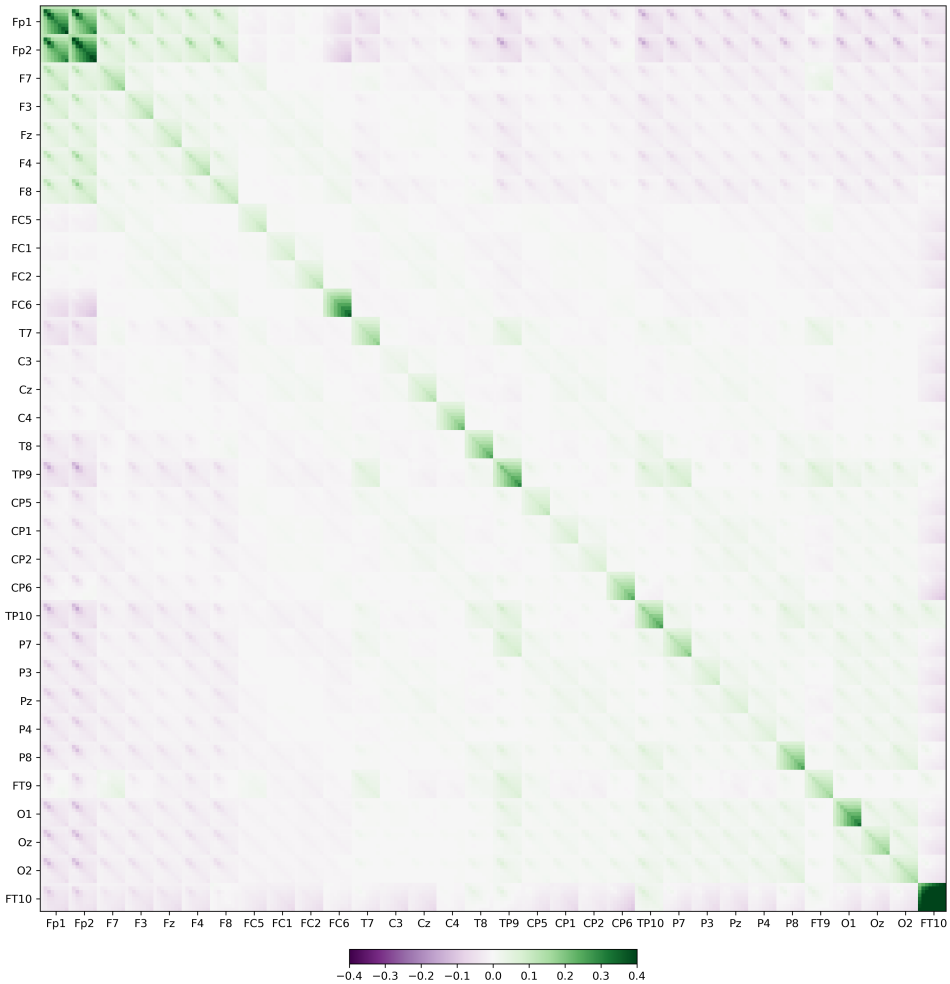


Figure 10: Average of per-participant covariance matrices (Σ).

6 Conclusions and discussion

In this thesis we studied the effects of word informativeness to brain activity. Additionally, our aim was to predict word informativeness from brain signals of persons reading natural language text.

6.1 Contributions

According to significance tests conducted with linear mixture models, word informativeness had a significant effect on ERP-components P200, P300, and P600 ($p < 0.05$). Furthermore, we were able to predict the informativeness of words based on the ERPs associated with them using single-trial LDA classifiers. The average AUC score of the classifiers was 0.643, and the classifiers performed better than random on all of the subjects ($p < 0.01$).

In the light of these results, we conclude that word informativeness, estimated statistically from natural text, has an effect on brain activity associated with reading (H1). In other words, the cognitive processing of informative words differs from that of uninformative words. The significant correlations with word informativeness and the P300 and P600 ERP-components are in line with the context-updating theory of the P300 and the P600-as-P300 theory: upon encountering an informative word in a stream of words, the mental representation of the text gets updated to represent its topic. The significant correlation in the P200 component is harder to explain, as there exists a wide array of factors that affect it. Its early onset time hints at early sensory processing, and indeed the P200 component is affected by the physical attributes of the stimulus [57]. However, the P200 has also been found to index higher cognitive processes, such as memory [13] and attention [53]. According to a prominent theory, the P200 indexes repetition suppression, a reduction in neural activity when a stimulus is repeated [25, 20]. The repetition suppression theory could explain the lower amplitude of the P200 component for uninformative words, as they are repeated multiple times in the documents.

Moreover, as shown by the classification performance scores and illustrated by the predicted words in Table 4, word informativeness can be predicted from event-related potentials in a single-trial setup (H2). Furthermore, the experiments were conducted in such a way that theoretically it would be possible to predict the informativeness on-line, and the predictions could possibly be utilized in a passive BCI system in real time.

6.2 Limitations

Due to the on-line assumption, we had made some amends in data preprocessing. Manual methods for preprocessing, such as ICA were not used, which might cause the data to be noisier. In general, however, the data preprocessing was conducted according to common standards [48], and should ensure that most of the artefacts caused by typical noise sources were removed. Producing two datasets with differing preprocessing methods for the neurophysiological experiment and prediction of ERPs was considered as an option, but we wanted to use the same dataset for both experiments in order to not obfuscate the results. Furthermore, the results of the neurophysiological experiment were used to define the features for the classifiers. Differing datasets in the two experiments could have resulted in suboptimal choices in feature engineering.

While we chose to use linear mixed models to avoid statistical methods whose assumptions could not be met, not all of the factors causing non-independencies in the measurements could be included as random effects in the LMMs. This is mainly due to the fact that when studying natural language, there is little control over the stimuli presented to subjects. Factors such as sentence length, context of a word, and individual world knowledge

of subjects amongst others may or may not affect the results. Including all of the factors possibly affecting the results as random effects in an LMM is infeasible for two reasons: not all of such factors may be known, and, on a more technical note, an LMM will fail to converge with too many factors. This failure to converge is caused by the model running out of degrees of freedom due to constraints imposed on the data by fixed and random effects. For the aforementioned reasons, we picked the factors that we deemed caused most of the variance in the measurements as effects in the LMMs.

Word length was found to be correlated with informativeness. Since studying the effect of word length to brain activity was not one of the goals of this thesis, we included word length as a random factor in the LMMs to mitigate its effect on the results. However, it is not clear whether the effect of word length can be separated from that of informativeness. Another question is, should the effect of word length be separated from that of informativeness in the first place? The correlation between word length and informativeness may be a sign that language is simply structured in such a way that shorter words are less informative than longer words. This idea has been presented by Piantadosi et al. in their study on efficient communication [61].

The classification results may have been improved by using a classifier other than LDA. LDA was chosen as the classifier mainly because it has been shown to perform well in single-trial ERP studies [9, 52, 46], and because the features learned by it could give insight on neural phenomena and noise sources in the data [31]. Furthermore, the focus of this work was not to present a high-performance classifier for ERPs, but to show that the prediction of word informativeness from ERPs was possible.

6.3 Implications

In this work we have demonstrated, to our knowledge for the first time, that informativeness of words is reflected in human cognitive processing of language. This suggests that the linguistic processing in the brain reflects the statistical nature of language. We hope this study paves way for further research uncovering the relationship between human cognition and language, and rouses interest in studying language outside the traditional scope of syntax and semantics. Future work could investigate the correlations of human linguistic processing with relation to other computational models of language, which would benefit both the neurolinguistic as well as BCI communities.

Acknowledgements

Special thanks to Paula Bergman, who shared her expertise on statistical models asking for nothing in return. The author would also like to thank the CNB study group for supporting him during his student years at the University of Helsinki.

References

[1] Adrian, E. D., and Matthews, B. H. The Berger rhythm: potential changes from the occipital lobes in man. *Brain 57*, 4 (1934), 355–385.

[2] Amzica, F., and Steriade, M. Electrophysiological correlates of sleep delta waves. *Electroencephalography and Clinical Neurophysiology 107*, 2 (Apr. 1998), 69–83.

[3] Banquet, J. P. Spectral analysis of the EEG in meditation. *Electroencephalography and Clinical Neurophysiology 35*, 2 (Aug. 1973), 143–151.

[4] Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language 68*, 3 (Apr. 2013), 255–278.

[5] Bashore, T. R., and van der Molen, M. W. Discovery of the P300: A tribute. *Biological Psychology 32*, 2 (Oct. 1991), 155–171.

[6] Bladin, P. F. W. Grey Walter, pioneer in the electroencephalogram, robotics, cybernetics, artificial intelligence. *Journal of Clinical Neuroscience 13*, 2 (Feb. 2006), 170–177.

[7] Blankertz, B., Curio, G., and Müller, K.-R. Classifying Single Trial EEG: Towards Brain Computer Interfacing. *Advances in neural information processing systems 14* (2002), 157–164.

[8] Blankertz, B., Dornhege, G., Krauledat, M., Tangermann, M., Williamson, J., Murray-Smith, R., and Müller, K.-R. The Berlin brain-computer interface presents the novel mental typewriter Hex-O-Spell. *Clinical Neurophysiology 113* (Jan. 2006).

[9] Blankertz, B., Lemm, S., Treder, M., Haufe, S., and Müller, K.-R. Single-trial analysis and classification of ERP components — A tutorial. *NeuroImage 56*, 2 (May 2011), 814–825.

[10] Boisgontier, M. P., and Cheval, B. The anova to mixed model transition. *Neuroscience & Biobehavioral Reviews 68* (Sept. 2016), 1004–1005.

[11] Brouwer, H., Fitz, H., and Hoeks, J. Getting real about Semantic Illusions: Rethinking the functional role of the P600 in language comprehension. *Brain Research 1446* (Mar. 2012), 127–143.

[12] Coenen, A., Fine, E., and Zayachkivska, O. Adolf Beck: A Forgotten Pioneer in Electroencephalography. *Journal of the History of the Neurosciences 23*, 3 (July 2014), 276–286.

[13] Dunn, B. R., Dunn, D. A., Languis, M., and Andrews, D. The Relation of ERP Components to Complex Memory Processing. *Brain and Cognition 36*, 3 (Apr. 1998), 355–376.

[14] Dustman, R. E., Boswell, R. S., and Porter, P. B. Beta Brain Waves as an Index of Alertness. *Science 137*, 3529 (Aug. 1962), 533–534.

[15] Eling, P. *Reader in the History of Aphasia : From (Franz) Gall to (Norman) Geschwind*. Amsterdam Studies in the Theory and History of Linguistic Science. Series II, Classics in Psycholinguistics. John Benjamins Publishing Company, Amsterdam, 1994.

[16] Eugster, M. J. A., Ruotsalo, T., Spapé, M. M., Barral, O., Ravaja, N., Jacucci, G., and Kaski, S. Natural brain-information interfaces: Recommending information by relevance inferred from human brain signals. *Scientific Reports 6* (Dec. 2016), 38580.

[17] Farwell, L. A., and Donchin, E. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology 70*, 6 (Dec. 1988), 510–523.

[18] Fedorenko, E., Scott, T. L., Brunner, P., Coon, W. G., Pritchett, B., Schalk, G., and Kanwisher, N. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences 113*, 41 (Oct. 2016), E6256–E6262.

[19] Fink, A., Grabner, R. H., Benedek, M., Reishofer, G., Hauswirth, V., Fally, M., Neuper, C., Ebner, F., and Neubauer, A. C. The creative brain: Investigation of brain activity during creative problem solving by means of EEG and FMRI. *Human Brain Mapping 30*, 3 (Mar. 2009), 734–748.

[20] Freunberger, R., Klimesch, W., Doppelmayr, M., and Höller, Y. Visual P2 component is related to theta phase-locking. *Neuroscience Letters 426*, 3 (Oct. 2007), 181–186.

[21] Friederici, A. D., Pfeifer, E., and Hahne, A. Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive brain research 1*, 3 (1993), 183–192.

[22] Friedman, D., Simson, R., Ritter, W., and Rapin, I. The late positive component (P300) and information processing in sentences. *Electroencephalography and Clinical Neurophysiology 38*, 3 (Mar. 1975), 255–262.

[23] Friedman, J. H. Regularized Discriminant Analysis. *Journal of the American Statistical Association 84*, 405 (Mar. 1989), 165–175.

[24] Frisch, S., Schlesewsky, M., Saddy, D., and Alpermann, A. The P600 as an indicator of syntactic ambiguity. *Cognition 85*, 3 (2002), B83 – B92.

[25] Golob, E. J., Pratt, H., and Starr, A. Preparatory slow potentials and event-related potentials in an auditory cued attention task. *Clinical Neurophysiology 113*, 10 (Oct. 2002), 1544–1557.

[26] Gouvea, A. C., Phillips, C., Kazanina, N., and Poeppel, D. The linguistic processes underlying the P600. *Language and Cognitive Processes 25*, 2 (Feb. 2010), 149–188.

[27] Haas, L. F. Hans Berger (1873–1941), Richard Caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry 74*, 1 (Jan. 2003), 9–9.

[28] Hagoort, P. Interplay between Syntax and Semantics during Sentence Comprehension: ERP Effects of Combining Syntactic and Semantic Violations. *Journal of Cognitive Neuroscience 15*, 6 (Aug. 2003), 883–899.

[29] Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K. M. Integration of Word Meaning and World Knowledge in Language Comprehension. *Science 304*, 5669 (Apr. 2004), 438–441.

[30] Hagoort, P., and Kutas, M. Electrophysiological insights into language deficits. In *Handbook of neuropsychology: Vol. 10 (pp. 105-134). Elsevier.*, vol. 10. Elsevier, 1995, pp. 105–134.

[31] Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., and Biessmann, F. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage 87* (Feb. 2014), 96–110.

[32] Herten, M. v., Kolk, H. H. J., and Chwilla, D. J. An ERP study of P600 effects elicited by semantic anomalies. *Cognitive Brain Research 22*, 2 (2005), 241 – 255.

[33] Isreal, J. B., Chesney, G. L., Wickens, C. D., and Donchin, E. P300 and Tracking Difficulty: Evidence For Multiple Resources in Dual-Task Performance. *Psychophysiology 17*, 3 (1980), 259–273.

[34] Kappenman, E. S., and Luck, S. J., Eds. *The Oxford Handbook of Event-Related Potential Components*, 1 ed. Oxford University Press, Dec. 2011.

[35] Katamba, F. What is a word? In *English words: structure, history, usage*. Routledge, 2015, p. 11. OCLC: 932062484.

[36] Klem, G. H., Lüders, H. O., Jasper, H. H., and Elger, C. The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalography and clinical neurophysiology. Supplement 52* (1999), 3–6.

[37] Kutas, M., and Federmeier, K. D. Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology 62*, 1 (Dec. 2010), 621–647.

[38] Kutas, M., and Hillyard, S. A. Reading senseless sentences: brain potentials reflect semantic incongruity. *Science 207*, 4427 (Jan. 1980), 203–205.

[39] Kutas, M., and Hillyard, S. A. Brain potentials during reading reflect word expectancy and semantic association. *Nature 307*, 5947 (Jan. 1984), 161–163.

[40] Kutas, M., and Van Petten, C. Event-Related Brain Potential Studies of language. In *Advances in psychophysiology*, vol. 3. JAI Press, Greenwich, Conn., 1988, pp. 139–187. OCLC: 959835982.

[41] Kwak, N.-S., Müller, K.-R., and Lee, S.-W. A convolutional neural network for steady state visual evoked potential classification under ambulatory environment. *PLOS ONE 12*, 2 (Feb. 2017), e0172578.

[42] Kübler, A., Kotchoubey, B., Kaiser, J., Wolpaw, J. R., and Birbaumer, N. Brain-computer communication: Unlocking the locked in. *Psychological Bulletin 127*, 3 (2001), 358–375.

[43] Laar, B. v. d., Gürkök, H., Bos, D. P.-O., Poel, M., and Nijholt, A. Experiencing BCI Control in a Popular Computer Game. *IEEE Transactions on Computational Intelligence and AI in Games 5*, 2 (June 2013), 176–184.

[44] Lau, E. F., Phillips, C., and Poeppel, D. A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience 9* (Dec. 2008), 920 EP –.

[45] Ledoit, O., and Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*, 2 (Feb. 2004), 365–411.

[46] Lemm, S., Blankertz, B., Curio, G., and Muller, K. Spatio-spectral filters for improving the classification of single trial EEG. *IEEE Transactions on Biomedical Engineering 52*, 9 (Sept. 2005), 1541–1548.

[47] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering 4*, 2 (June 2007), R1–R13.

[48] Luck, S. J. Artifact Rejection and Correction. In *An introduction to the event-related potential technique*, 2nd ed. MIT Press, Cambridge, Massachusetts, 2014, pp. 185–217.

[49] Luck, S. J. *An introduction to the event-related potential technique*, 2nd ed. The MIT Press, Cambridge, Massachusetts, 2014.

[50] Makeig, S., Bell, A. J., Jung, T.-P., and Sejnowski, T. J. Independent Component Analysis of Electroencephalographic Data. *Advances in Neural Information Processing Systems 8* (1996), 145–151.

[51] Mu, Y., Kitayama, S., Han, S., and Gelfand, M. J. How culture gets embrained: Cultural differences in event-related potentials of social norm violations. *Proceedings of the National Academy of Sciences 112*, 50 (Dec. 2015), 15348–15353.

[52] Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. Machine learning for real-time single-trial EEG-analysis: From brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods 167*, 1 (Jan. 2008), 82–90.

[53] Neville, H. J., and Lawson, D. Attention to central and peripheral visual space in a movement detection task: an event-related potential and behavioral study. I. Normal hearing adults. *Brain Research 405*, 2 (Mar. 1987), 253–267.

[54] Nunez, P. L. *Electric fields of the brain : the neurophysics of EEG*, 2nd ed. Oxford University Press, New York, 2006.

[55] Obler, L. K. *Language and the brain.* Cambridge approaches to linguistics. Cambridge University Press, Cambridge, 1999.

[56] Ojala, M., and Garriga, G. C. Permutation Tests for Studying Classifier Performance. In *2009 Ninth IEEE International Conference on Data Mining* (Miami Beach, FL, USA, Dec. 2009), IEEE, pp. 908–913.

[57] Omoto, S., Kuroiwa, Y., Otsuka, S., Baba, Y., Wang, C., Li, M., Mizuki, N., Ueda, N., Koyano, S., and Suzuki, Y. P1 and P2 components of human visual evoked potentials are modulated by depth perception of 3-dimensional images. *Clinical Neurophysiology 121*, 3 (Mar. 2010), 386–391.

[58] Osterhout, L., and Holcomb, P. J. Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language 31*, 6 (Dec. 1992), 785–806.

[59] Perlis, M. L., Merica, H., Smith, M. T., and Giles, D. E. Beta EEG activity and insomnia. *Sleep Medicine Reviews 5*, 5 (Oct. 2001), 365–376.

[60] Perrin, F., Pernier, J., Bertrand, O., and Echallier, J. F. Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology 72*, 2 (1989), 184 – 187.

[61] Piantadosi, S. T., Tily, H., and Gibson, E. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences 108*, 9 (Mar. 2011), 3526–3529.

[62] Polich, J. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology 118*, 10 (2007), 2128 – 2148.

[63] Ponte, J. M., and Croft, W. B. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New York, NY, USA, 1998), SIGIR '98, ACM, pp. 275–281.

[64] Porter, M. An algorithm for suffix stripping. *Program 14*, 3 (Mar. 1980), 130–137.

[65] Sassenhagen, J., Schlesewsky, M., and Bornkessel-Schlesewsky, I. The P600-as-P3 hypothesis revisited: Single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain and Language 137* (Oct. 2014), 29–39.

[66] Smith, S. J. M. EEG in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry 76*, Suppl II (June 2005), ii2–ii7.

[67] Song, F., and Croft, W. B. A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management - CIKM '99* (Kansas City, Missouri, United States, 1999), ACM Press, pp. 316–321.

[68] Sutton, S., Braren, M., Zubin, J., and John, E. R. Evoked-Potential Correlates of Stimulus Uncertainty. *Science 150*, 3700 (Nov. 1965), 1187–1188.

[69] Tanner, D., Morgan-Short, K., and Luck, S. J. How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition: High-pass filtering and artifactual ERP effects. *Psychophysiology 52*, 8 (Aug. 2015), 997–1009.

[70] Urigüen, J., and Zapirain, B. EEG artifact removal – State-of-the-art and guidelines. *Journal of neural engineering 12* (Apr. 2015), 031001.

[71] Vanderwolf, C. H. Are neocortical gamma waves related to consciousness? *Brain Research 855*, 2 (Feb. 2000), 217–224.

[72] Vidal, J. J. Toward Direct Brain-Computer Communication. *Annual Review of Biophysics and Bioengineering 2*, 1 (1973), 157–180.

[73] Vidal, J. J. Real-time detection of brain events in EEG. *Proceedings of the IEEE 65*, 5 (May 1977), 633–641.

[74] Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. Brain–computer interfaces for communication and control. *Clinical Neurophysiology 113*, 6 (June 2002), 767–791.

[75] Wolpaw, J. R., McFarland, D. J., Neat, G. W., and Forneris, C. A. An EEG-based brain-computer interface for cursor control. *Electroencephalography and Clinical Neurophysiology 78*, 3 (Mar. 1991), 252–259.

[76] Xue, J., and Hall, P. Why Does Rebalancing Class-Unbalanced Data Improve AUC for Linear Discriminant Analysis? *IEEE Transactions on Pattern Analysis and Machine Intelligence 37*, 5 (May 2015), 1109–1112.

[77] Xue, J.-H., and Titterington, D. M. Do unbalanced data have a negative effect on LDA? *Pattern Recognition 41*, 5 (May 2008), 1558–1571.

[78] Young, G. B. The EEG in Coma. *Journal of Clinical Neurophysiology 17*, 5 (Sept. 2000), 473.

[79] Zander, T. O., and Kothe, C. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. *Journal of Neural Engineering 8*, 2 (Apr. 2011), 025005.

[80] Zhai, C. Statistical Language Models for Information Retrieval. *Synthesis Lectures on Human Language Technologies 1*, 1 (Jan. 2008), 1–141.

[81] Zhai, C., and Lafferty, J. A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. *SIGIR Forum 51*, 2 (Aug. 2017), 268–276.

A EEG data measurement experiment details

The following text and figures A1 and A2 have been reproduced[1] from the Supplementary Information of [16], a previously conducted study. The text mentions the amplifier running at 200 Hz, while actually the amplifier ran at 2000 Hz as stated in the main matter of this thesis.

SI Neural Activity Recording Experiment

We recorded the electroencephalography (EEG) signals of 17 participants while each participant performed a set of eight reading tasks. The following sections provide the experimental details.

Participants

Participants were volunteers recruited from the universities of the Helsinki metropolitan area in Finland. They were selected only if they were right-handed, had no self-reported neuropathological history, and were deemed to have sufficient fluency in English. Handedness was assessed using the Edinburgh Handedness Inventory[1,2] and English fluency using the Cambridge English "Test your English - Adult Learners" online test[3]. Seventeen participants were recruited to participate in the experiment. The data of two participants were discarded due to technical issues. Of the fifteen remaining, 8 were female and 7 male. Their English fluency was assessed as high (Mean = 23.53, SD = 1.23; maximum value is 25), and their handedness as right-handed (Mean = 87.35, SD = 12.13; the index range between -100 which is fully left-handed to +100 which is fully right-handed). They were fully briefed as to the nature and purpose of the study prior to the experiment. Furthermore, and in accordance with the Declaration of Helsinki, they signed informed consents and were instructed on their rights as participants, including the right to withdraw from the experiment at any time without fear of negative consequences. They received two movie tickets as compensation for their participation.

Procedure and Design

Following the initial briefing, participants were explained the task in more detail, while the EEG equipment was set up. They then received a short training task with two sample topics. When participants indicate their complete understanding of the task, the experiment commenced. One experimental block was called a reading task. During one reading task two documents were read. The two documents were randomly drawn (without replacement) from the pool of 30 document candidates. A document was defined as the first six sentences of the corresponding Wikipedia article. Participants completed eight reading task.

[1]Used under CC BY 4.0 license https://creativecommons.org/licenses/by/4.0/. The figure and citation numbering were changed, otherwise no changes were made to the original.

Figure A1 shows the step-by-step explanation of a reading task (also called a block in the visualization) each participant received during the initial briefing. At the beginning of one reading task participants were asked to freely choose which one of the two documents should be the relevant topic and which one the irrelevant topic. Every reading task comprised six trials, each consisting of one sentence form the relevant and one sentence from the irrelevant document. Each trial consisted of the sequential presentation of words (the word stream), which the participants should "just read" and during which their brain signals were monitored. After that, two validity sub-tasks had to be fulfilled to ensure the active participation of the participants. Finally, in the explicit word relevance judgment task, the participants rated the "just read" words as "relevant" or "irrelevant". The explicit judgment task provided the labels for the data analysis. This was needed as we were interested in the subjective relevance of each participant.

Figure A2 shows the concrete implementation of a reading task as a cognitive neuroscience experiment. Every trial started with a warning signal (the words "Starting trial"), followed by the presentation of the mask. An initial sentence separator (a randomized sequence of 4 to 9 numbers or other non-alphabetic characters like `%&$`) was shown before the word stream was shown. The word stream consisted of the sequential presentation of each word in the first sentence, followed by a sentence separator, the words in the second sentence, and concluded by a final sentence separator. Every word and sentence separator was presented for exact 699 ms (SD = 0.3 ms). Punctuation marks were not shown. Masking effects were countered to some extent by the frame resizing, which keeps the level of foveal stimulation constant. In our previous experiment on term-relevance prediction[6] and during the pilot experiments for this experiment, we learned that people had more difficulty reading with than without short masks between the bursts, so as a consequence we removed them. It is possible these masking effects may be much more significant with strong "flashing", as would be the case with very short stimulus durations. Here, the words appearing at a slow rate of ca 700 ms per words. This reading pace was determined in our previous experiment on term-relevance prediction[6] and during the pilot experiments for this experiment. The reading pace was a compromise between being slow enough that the brain signals of two consecutive words to not overlap too much, and still being fast enough that a (more or less) fluent reading is possible.

Following the word stream, two extra sub-tasks were presented to validate that the participants had remembered their chosen word and that they had paid attention to both sentences. First, they were asked to type in the name of the relevant topic in order to ascertain they had not forgotten. Then, a recall task was presented to prevent the participants from selectively concentrating on one of the two sentences. One of the sentences was selected randomly and presented in full on the screen, with one of the nouns or verbs

substituted by question marks. Participants were asked to type in the word missing in the sentence. They were then presented with feedback in points regarding their performance on these two tasks as a motivational instrument (similar to [7]).

Then, in the final part of the trial, the participants were asked to explicitly rate the relevance of all words from the relevant topic. All words were shown in one (if the sentence comprised fewer than 35 words) or two columns on the screen. A cursor was presented next to each word, indicating a two-alternative forced-choice decision. Pressing the left arrow key on the keyboard would rate the word as irrelevant and pressing the right would rate it as relevant. Participants were instructed prior to the experiment that they should not re-interpret the relevance of the words and instead make a decision based on their previous viewing of the sentence. To facilitate this, they received a maximum of 2 s to respond to each word, after which the cursor moved to the next word in the sentence. After the last word was rated, the trial was completed, with the next trial starting after an inter-trial interval of ca. 1 s, unless it was the last trial in the block. After completing a block, they were requested to freely write about their chosen, relevant topic; this task was defined to keep the participant engaged. Finally, they filled out a questionnaire with two items for both topics, one regarding their interest ("how interesting do you find topic x") and one regarding their knowledge ("how much do you know about topic x") using a 9-point rating scale (1: not at all – 9: extremely so). Three self-timed breaks with a minimum of one minute evenly split the blocks into four parts. The experiment, excluding preparation and instruction, lasted approximately one hour.

Apparatus and Stimuli

Words were presented with an 18-point Lucida Console black typeface at the center of the 19" LCD screen. They were shown against a silver (RGB 82%, 82%, 82%) background in the middle of a 300 x 100 pixel pattern mask. The mask was a black rectangle with a grid-like pattern, with an opening to show the word. This was used to control the degree to which word length affected light reaching the eyes (i.e. to make sure longer words were not tantamount to more black pixels on the screen). Sentence separators were word-like character repetitions consisting of 4 to 9 numbers (`3333333`) or other non-alphabetic characters (`&&&&&&&`), which were designed to mimic the same early visual activity as words without evoking psycholinguistic processing.

The screen was positioned approximately 60 cm from the participants and was running at a resolution of 1680 x 1050 and a refresh rate of 60 Hz. Stimulus presentation, timing, and EEG synchronization were controlled using E-Prime 2 Professional 2.0.10.353 on a PC running Windows XP SP3. EEG was recorded from 32 Ag/AgCl electrodes, positioned on standardized (using EasyCap elastic caps, EasyCap GmbH, Herrsching, Germany), equidis-

tant electrode sites of the 10 - 20 system via a QuickAmp (BrainProducts GmbH, Gilching, Germany) amplifier running at 200 Hz. Additionally, the electro-oculogram for vertical eye movements (and eye blinks) and horizontal eye movements was recorded using bipolar electrodes positioned respectively 2 cm superior/inferior to the right pupil and 1 cm lateral to the outer canthi of both eyes.

Pilot experiments

Prelimary versions of the final experimental procedure and design were piloted with four separate participants. In these experiments, we tested and evaluated, for example, the stimulus duration, the explicit feedback task, and the points system. The data of these pilot experiments were not used in the final analysis, except that some basic parameter estimations for the final feature engineering process were based on cross-validation experiments on these data (e.g., number of feature windows).

References

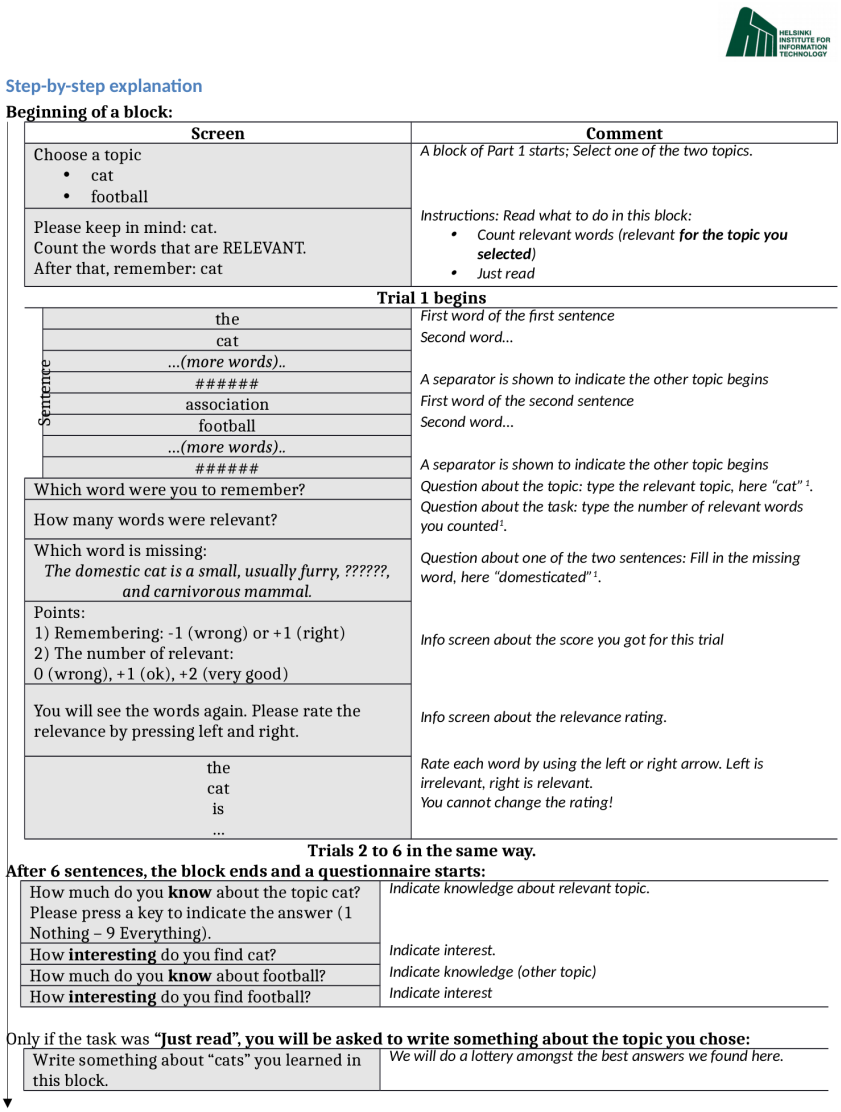
1. Oldfield, O. R. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia 9, 97–113 (1971).

2. Cohen, M. S. Handedness questionnaire (2014). http://www.brainmapping.org/shared/Edinburgh.php.

3. Cambridge English Language Assessment. Test your English – Adult Learners (2014). http://www.cambridgeenglish.org/test-your-english/adult-learners/.

4. Eugster, M. J. A. et al. Predicting term-relevance from brain signals. In Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, 425-434 (ACM, New York, NY, USA, 2014).

5. Spape, M. M., Band, G. P. & Hommel, B. Compatibility-sequence effects in the Simon task reflect episodic retrieval but not conflict adaptation: Evidence from LRP and N2. Biological psychology 88, 116-123 (2011).



Step-by-step explanation

Beginning of a block:

Screen	Comment
Choose a topic • cat • football	*A block of Part 1 starts; Select one of the two topics.*
Please keep in mind: cat. Count the words that are RELEVANT. After that, remember: cat	*Instructions: Read what to do in this block:* • *Count relevant words (relevant **for the topic you selected**)* • *Just read*
Trial 1 begins	
Sentence: the	*First word of the first sentence*
Sentence: cat	*Second word…*
Sentence: *…(more words)..*	
Sentence: ######	*A separator is shown to indicate the other topic begins*
Sentence: association	*First word of the second sentence*
Sentence: football	*Second word…*
Sentence: *…(more words)..*	
Sentence: ######	*A separator is shown to indicate the other topic begins*
Which word were you to remember?	*Question about the topic: type the relevant topic, here "cat" [1].*
How many words were relevant?	*Question about the task: type the number of relevant words you counted[1].*
Which word is missing: *The domestic cat is a small, usually furry, ??????, and carnivorous mammal.*	*Question about one of the two sentences: Fill in the missing word, here "domesticated" [1].*
Points: 1) Remembering: -1 (wrong) or +1 (right) 2) The number of relevant: 0 (wrong), +1 (ok), +2 (very good)	*Info screen about the score you got for this trial*
You will see the words again. Please rate the relevance by pressing left and right.	*Info screen about the relevance rating.*
the cat is …	*Rate each word by using the left or right arrow. Left is irrelevant, right is relevant.* *You cannot change the rating!*
Trials 2 to 6 in the same way.	

After 6 sentences, the block ends and a questionnaire starts:

How much do you **know** about the topic cat? Please press a key to indicate the answer (1 Nothing – 9 Everything).	*Indicate knowledge about relevant topic.*
How **interesting** do you find cat?	*Indicate interest.*
How much do you **know** about football?	*Indicate knowledge (other topic)*
How **interesting** do you find football?	*Indicate interest*

Only if the task was **"Just read", you will be asked to write something about the topic you chose:**

Write something about "cats" you learned in this block.	*We will do a lottery amongst the best answers we found here.*

1 Answer as fast as possible!

Figure A1: Step-by-step explanation of the experiment. The illustrations shows the composition of one reading task (block). A participant conducted eight such reading tasks with each one with different documents. This explanation was part of the information sheet all participants received during their initial briefing.

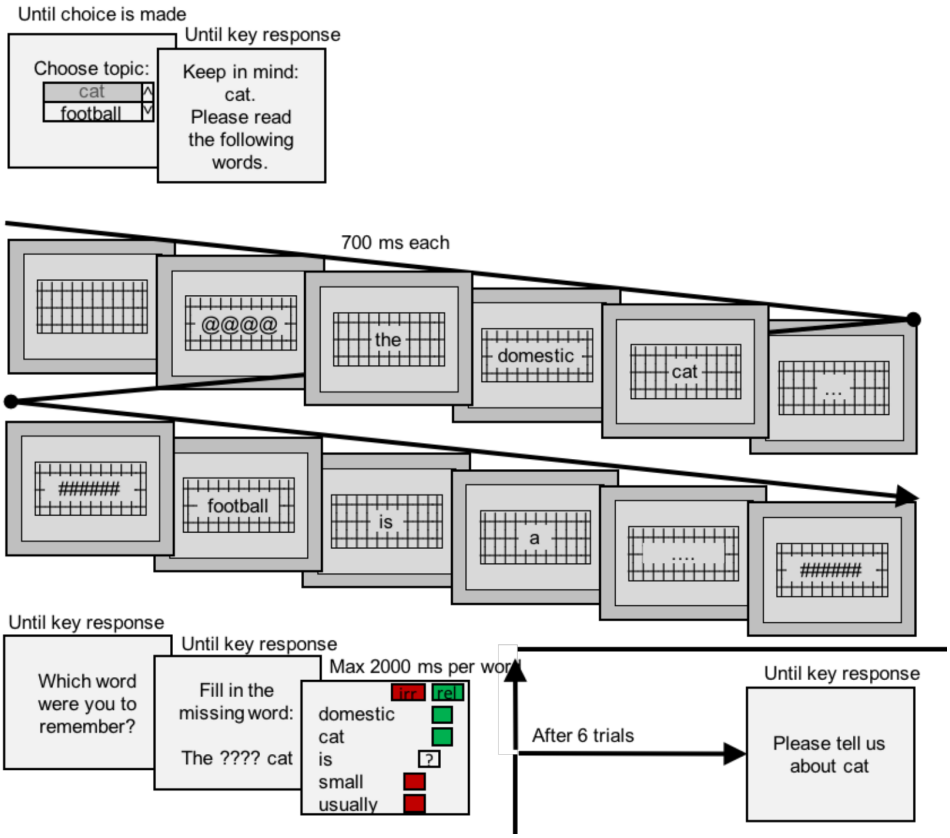


Figure A2: Illustration of the technical implementation as a cognitive neuroscience experiment. The figure shows the screen by screen execution of the block described in Figure A1.

B Preprocessing details

Subject	Threshold (μV)	Epochs recorded	Epochs dropped	Channels dropped
S01	57.42	1941	388	None
S02	33.88	1961	392	Fp1, Fp2, TP9, TP10, FT10
S03	65.54	1936	387	Fp1, Fp2
S04	30.64	1986	397	Fp1, Fp2, P7
S05	31.19	1959	391	Fp1, Fp2, F7, TP9, TP10
S06	51.04	1960	392	Fp1, Fp2, O2
S07	27.98	1869	373	TP10
S08	62.90	1958	391	Fp1, Fp2, TP9
S09	47.25	1818	363	None
S10	28.69	2026	405	Fp1, Fp2, O2
S11	57.04	1939	387	None
S12	40.61	1944	388	Fp1, Fp2, F7, TP9
S13	35.28	1869	379	Fp1, Fp2
S14	29.96	1981	396	Fp1, Fp2, F7, FT9, FT10
S15	44.96	1906	381	Fp1, Fp2, F7

Table B1: EEG preprocessing details.

C EEG visualizations

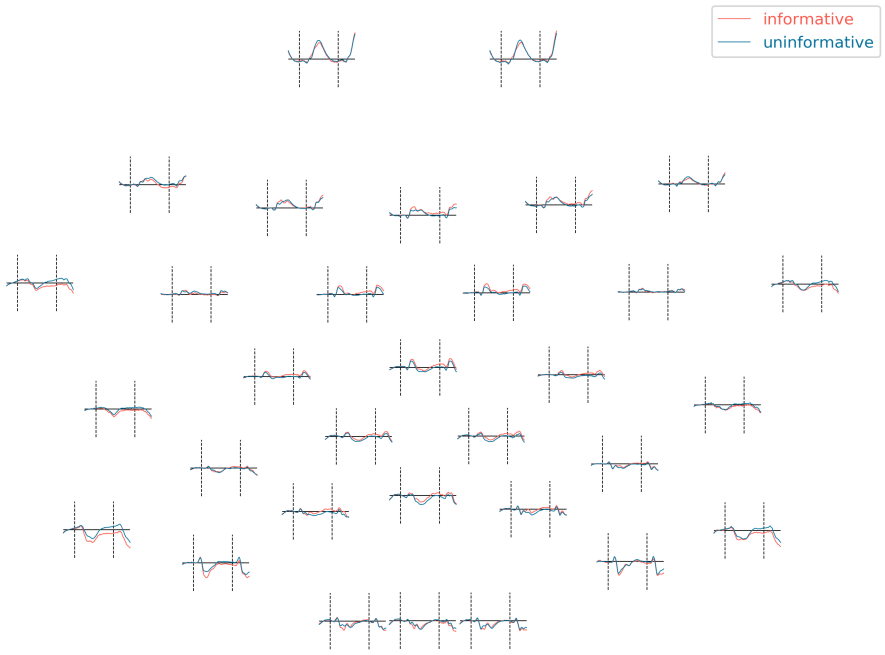


Figure C1: ERPs for all channels.

D Word classifications for all documents

	Top/bottom 5 words in informativeness class:			
Document topic	Informative predicted	Informative true	Uninformative predicted	Uninformative true
atom	positively	protons	such	as
	successfully	decay	or	is
	principles	isotope	one	and
	neutrons	neutrons	be	to
	quantum	atoms	and	the
automobile	primarily	benz	or	as
	affordable	motorcar	after	is
	denote	automobile	than	in
	automobile	automobiles	soon	and
	motorized	risen	when	to
bank	liabilities	paschi	on	as
	importance	berenberg	into	is
	markets	institutionalised	are	in
	renaissance	siena	either	and
	activities	intermediary	those	to
bicycle	automobiles	sprockets	to	as
	bike	bicyclist	has	is
	automobile	bicycle	around	in
	played	bicycles	on	and
	eventually	cyclist	ball	to
bill clinton	democrat	clinton	over	as
	arkansas	arkansas	an	is
	presided	peacetime	generation	in
	foundation	boomer	who	and
	agreement	42nd	in	to
brain	generating	synapses	with	as
	organ	invertebrate	balance	is
	vertebrate	neurons	view	in
	special	cortex	each	and
	primary	cerebral	for	to
cat	mammal	felids	as	they
	housecat	housecat	and	as
	indoor	felines	no	is
	despite	cats	in	in
	flexible	cat	too	and
communism	distinction	marxism	absence	as
	marxism	dictatorship	in	is
	maximized	marx	for	in
	influenced	communist	today	and
	significantly	automation	and	to

euro	following	eurozone	2002	as
	dollar	euro	july	is
	eurozone	banking	after	in
	debt	currency	as	and
	adopted	banknotes	is	the
football	penalty	soccer	as	as
	spherical	ball	are	is
	opposing	football	into	in
	england	torso	use	and
	touch	codified	body	to
india	asia	indus	in	is
	civilisation	multilingual	vast	in
	independent	gandhi	to	and
	subcontinent	mahatma	led	to
	mahatma	civilisation	of	the
learning	consciously	habituation	of	as
	machines	factual	in	is
	conscious	machines	by	in
	synthesizing	learning	more	and
	intelligent	learn	involve	to
machine learning	algorithm	unsupervised	by	is
	filtering	subfield	search	in
	unsupervised	algorithm	is	and
	outputs	spam	with	to
	subfield	inputs	it	the
michael jackson	brothers	moonwalk	to	as
	publicized	jackson	him	is
	techniques	thriller	such	in
	appearance	michael	an	and
	michael	robot	along	to
money	commodity	fiat	it	as
	repayment	deferred	an	is
	market	commodity	to	in
	coins	tender	unit	and
	emergent	bank	as	to
ocean	hydrosphere	hydrosphere	on	is
	conventional	hadean	which	in
	emergence	saline	in	and
	divisions	oceanographers	an	to
	believed	ocean	and	the
painting	applied	airbrushes	western	as
	surface	pigment	to	is
	spiritual	paint	trade	in
	dominated	painting	or	and
	builders	brush	but	to

plato	philosopher	socrates	an	as
	philosophical	socratic	in	is
	aristotle	plato	is	in
	dialogues	platonism	been	and
	athens	dialogues	the	to
politics	democracy	adversaries	in	is
	exercised	civic	one	in
	negotiation	discourse	from	and
	international	clans	or	to
	practice	warfare	among	the
rome	architecture	bramante	to	as
	bramante	tiber	in	is
	michelangelo	bernini	was	in
	resided	lazio	chapel	and
	oldest	michelangelo	is	to
savanna	majority	savannas	common	is
	seasonal	savanna	of	in
	confined	herbaceous	so	and
	sufficient	canopy	and	to
	herbaceous	grassland	a	the
schizophrenia	schizophrenia	antipsychotic	primarily	is
	environment	hallucinations	what	and
	characterized	schizophrenia	by	to
	additional	contributory	a	the
	psychological	symptoms	of	of
school	teenagers	homeschooling	is	as
	building	vocational	a	is
	compulsory	secondary	who	in
	economics	compulsory	to	and
	university	learning	be	to
society	insofar	subculture	are	as
	institutions	ant	used	is
	colony	insofar	on	in
	extensively	interpersonal	that	and
	organism	artificial	by	to
star	spectrum	luminosity	to	is
	neutron	brightest	a	and
	plasma	stellar	held	to
	primarily	constellations	space	the
	metallicity	neutron	composed	of
telephone	numeric	landline	two	they
	connected	keypad	in	as
	telecommunications	earphone	of	is
	transmissions	telephone	to	in
	initiate	microphone	a	and

time	fundamental	technologists	all	as
	performing	intervals	component	is
	religion	judgement	be	in
	occupied	astronomy	a	and
	scientists	sensation	from	to
volcano	converging	volcano	type	as
	eruption	volcanoes	can	is
	droplets	lava	not	in
	plumes	eruptions	on	and
	rupture	erupting	lower	to
wife	widow	marital	from	as
	separated	spouse	also	is
	heterosexual	heterosexual	of	in
	varies	husband	term	and
	cultures	widow	who	to
wine	thousands	wine	so	is
	chemical	grapes	lets	in
	egyptians	fermented	has	and
	religion	ferment	in	the
	production	yeasts	or	of

Table D1: Top 5 words per topic sorted by classifier confidence (predicted) for class membership (informative/uninformative) and by informativeness (true).